Ciencia e Ingeniería Neogranadina 2019 Vol. 30(1) enero - julio 2020 ■ ISSN: 0124-8170 · ISSN-e: 1909-7735 ■ pp. 75-88

Editorial Neogranadina

DOI: https://doi.org/10.18359/rcin.4236

Hierarchical Clustering for Detecting Anomalous Traffic Conditions in Power Substations

Erwin Alexander Leal Piedrahita[a]

Abstract: The IEC 61850 standard has contributed significantly to substation management and automation by incorporating the advantages of communications networks into the operation of power substations. However, this modernization process also involves new challenges in other areas. For example, in the field of security, several academic works have shown that the same attacks used in computer networks (DoS, sniffing, tampering, spoofing, among others) can also compromise the operation of a substation. This article evaluates the applicability of hierarchical clustering algorithms and statistical type descriptors (averages) to the identification of anomalous traffic patterns in communication networks for IEC 61850 power substations. The results obtained show that using a hierarchical algorithm with the Euclidean distance proximity criterion and simple link grouping method, a correct classification is achieved in the following operation scenarios: 1) normal traffic, 2) IED disconnection, 3) network discovery attack, 4) DoS attack, 5) IED spoofing attack, and 6) failure of the high voltage line. In addition, the descriptors used for classification proved equally effective with other unsupervised clustering techniques such as K-means (partitional-type clustering) or LAMDA (diffuse-type clustering).

Keywords: Clustering; Hierarchical; IEC 61850; power substation; traffic detection; unsupervised learning

Received: July 31st, 2019 **Accepted:** October 9th, 2019

Available online: July 15th, 2020

How to cite: E. A. Leal Piedrahita, "Hierarchical Clustering for Anomalous Traffic Conditions Detection in Power Substations", *Cien.Ing.Neogranadina*, vol. 30, no. 1, pp. 75-88, Nov. 2019.

a Universidad de Antioquia. E-mail erwin.leal@udea.edu.co.
ORCID: https://orcid.org/0000-0001-6757-2538

Agrupamiento jerárquico para la detección de condiciones de tráfico anómalo en subestaciones de energía

Resumen: el estándar IEC 61850 ha contribuido notablemente con el proceso de gestión y automatización de las subestaciones, al incorporar las ventajas de las redes de comunicaciones en la operación de las subestaciones de energía. Sin embargo, este proceso de modernización también involucra nuevos desafíos en otros campos. Por ejemplo, en el área de la seguridad, diversos trabajos académicos han puesto en evidencia que la operación de una subestación también puede ser comprometida por los mismos ataques utilizados en las redes de cómputo (DoS, *sniffing*, *tampering*, *spoofing,* entre otros). Este artículo evalúa la aplicabilidad de los algoritmos de agrupamiento no supervisado de tipo jerárquico y el uso de descriptores de tipo estadístico (promedios), en la identificación de patrones de tráfico anómalo en redes de comunicación para subestaciones eléctricas basadas en el estándar IEC 61850. Los resultados obtenidos demuestran que, utilizando un algoritmo jerárquico con criterio de proximidad distancia Euclidiana y método de agrupación vínculo simple, se logra una correcta clasificación de los siguientes escenarios de operación: 1) Tráfico normal, 2) Desconexión de dispositivo IED, 3) Ataque de descubrimiento de red, 4) Ataque de denegación de servicio, 5) Ataque de suplantación de IED y 6) Falló en la línea de alta tensión. Además, los descriptores utilizados para la clasificación demostraron ser robustos al lograrse idénticos resultados con otras técnicas de agrupamiento no supervisado de tipo particional como K-medias o de tipo difuso como LAMDA (Learning Algorithm Multivariable and Data Analysis).

Palabras clave: Jerárquico; agrupamiento; aprendizaje no supervisado; IEC 61850; detección de tráfico; subestación eléctrica

Introduction

Smart Grid is a concept that aims to provide mechanisms for energy generation and consumption in a more efficient and intelligent manner. This concept proposes the adoption of data network advantages for grid operation in the areas of control, communications and monitoring [1]. For this purpose, the modernization of the infrastructure that supports power generation, transmission, distribution, and consumption has caused the emergence, within the communications network, of a variety of IP compliant devices that are interconnected by an Ethernet technology-based network [2].

Although the communication networks of IEC 61850 modern electrical substations [3] provide better benefits than the communication networks of traditional substations, companies are being cautious with their implementation due to the vulnerabilities noted by various research articles [4]. For example, in [5] and [6] Denial of Service (DoS) attacks were implemented. Also, in [6] network traffic was intercepted (sniffing). Interception and modification of critical traffic (tampering) were attained in [7] and [8], while a spoofing attack was achieved in [9].

In this context, anomaly or intrusion detections within power substation communication networks have become an important research topic, as a consequence of the serious damage that failure may cause in this critical infrastructure. The majority of intrusion detection systems are focused on the detection of signatures (characteristic patterns associated with a particular intrusion or attack). However, for obvious reasons, this detection does not cover new types of attacks [10]. Hence, the study of unsupervised classification techniques that, through the wide recognition of normal network traffic, can allow the identification of possible abnormal states of operation is of special interest.

The main contribution of this paper is to determine the application of hierarchical clustering algorithms in the identification of anomalous operation scenarios, specifically in IEC 61850 power substation communication networks.

Conceptual framework

The main notions of IEC 61850 communication networks, as well as the operating fundamentals of hierarchical clustering techniques, are presented below.

Communication networks in power substations

In general, we can define a communications network of an automated power substation as a set of IP-supported devices, exchanging information via an Ethernet network that uses switches as interconnecting elements. This network is set up to ensure a communication platform supporting management, monitoring, synchronization, protection, control and sensing operations within power substations. Currently, the substation automation process is guided by IEC 61850 [3], which covers almost all aspects of a Substation Automation System (SAS). This standard provides recommendations to guarantee the interoperability of devices from different manufacturers. Also, the standard defines how management, control and protection, and measurement devices intercommunicate inside a substation. As shown in Fig. 1, the model proposed by IEC 61850 is hierarchical, in which three levels are identified: station, bay, and process, interconnected via the process bus and the station bus.

The process level is composed of actuators, measuring devices called MUs, Ethernet switches and yard equipment such as Current Transformers (CTs), Voltage Transformers (VTs), and breakers. At the bay level, we find protection and control IEDs (Intelligent Electronic Devices), while at the station level Ethernet switches and communications network management devices are located. IEC 61850 also defines four types of communication services in order to ensure the correct operation of the network (Table 1) [2].

Pursuant to IEC 61850-5 and IEC 61850-8, it is recommended that communication services be mapped to different communication stacks according to their performance requirements (Fig. 2).

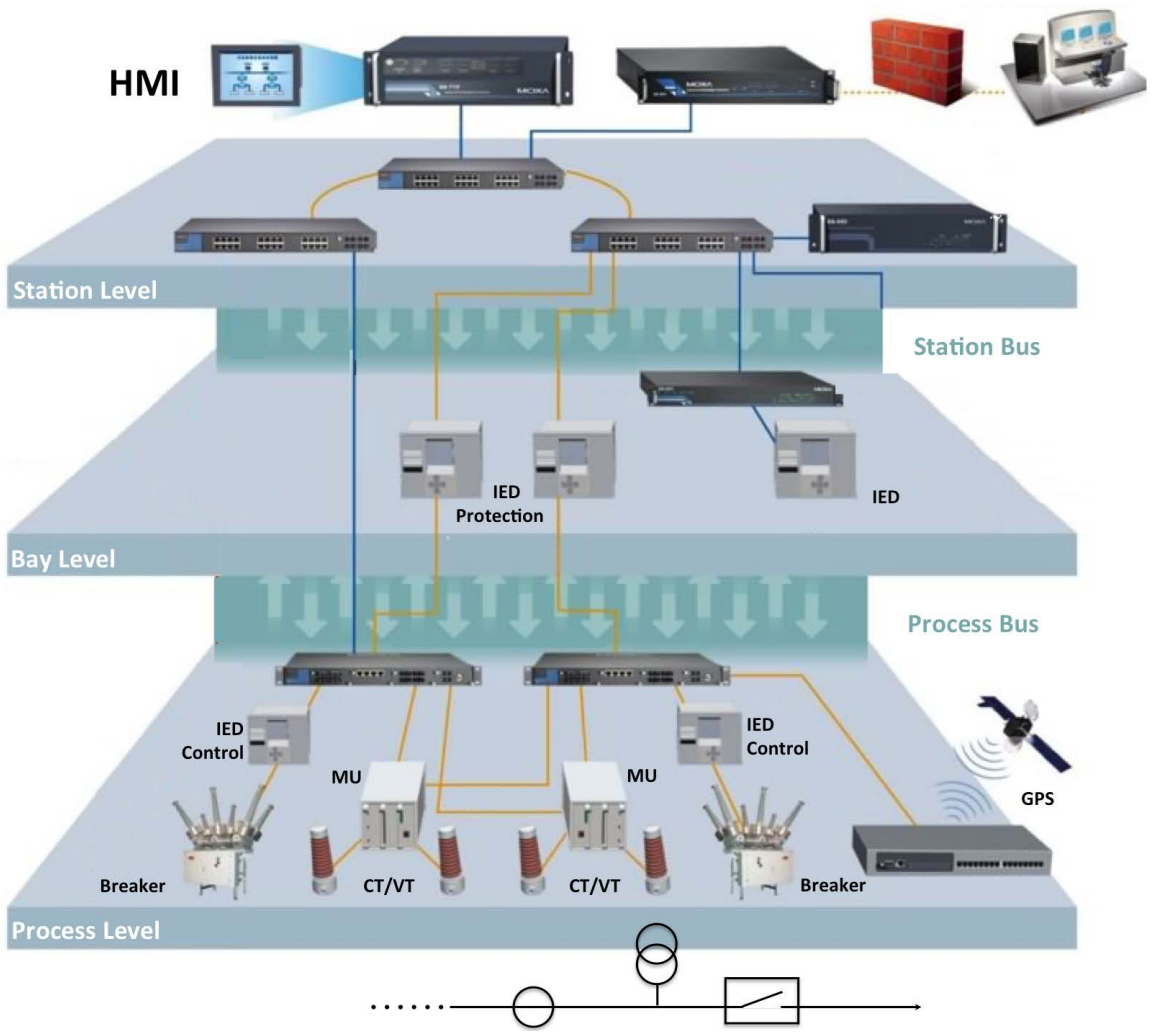

Fig. 1. Communication model for IEC 61850.

Source: The author.

Table 1. Communication services defined in IEC 61850

Type of Service	Description
ACSI (Abstract Communication Service Interface)	Defined in IEC 61850-7-2 to address the basic requirements for the process of exchanging information. With this aim, the MMS (Manufacturing Message Specification) protocol is used to transport operational information for managing the substation between the user interface system and IEDs.
GOOSE (Generic Object-Oriented Substation Event)	Defined in IEC 61850-8-1 for the purpose of distributing event data (commands, alarms, indications, trip messages) among IEDs across the entire substation network.
SMV / SV (Sampled Measured Values)	Specified in IEC 61850-9-2 and used to transmit analog values (current and voltage) from the MUs to the IEDs.
TS (Time Synchronization)	Uses the PTP (Precision Time Protocol) for ensuring clock synchronization among devices of a distributed system.

Source: The author.

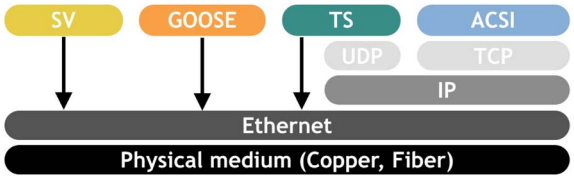

Fig. 2. Communication stack under IEC 61850.
Source: The author.

Hierarchical clustering algorithms

A clustering algorithm is a multivariate statistical procedure aiming at grouping or classifying the elements of a data space into compact, separate, and homogeneous groups called clusters or classes. In particular, unsupervised clustering algorithms intend to discover the composition of the classes or groupings to which elements belong without having prior information about data structure. This clustering should guarantee that the degree of natural association is high among members of the same group and low among members of diverse groups [11]. Unsupervised clustering algorithms are divided into two major categories: hierarchical and partitional. Partitional algorithms divide data space into a specified number of groups, following an optimization criterion. Meanwhile, hierarchical algorithms generate a structured organization of nested groups, which is represented by a classification tree known as a dendrogram (Fig. 3). The dendrogram illustrates how the algorithm groups elements step by step and, observing the structure of their branches and the distance among them, the diagram shows the degree of similarity between different clusters. In addition, depending on where the cut level of the dendrogram is established, the number of classes for the classification algorithm is defined [11].

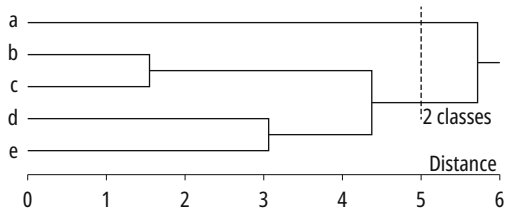

Fig. 3. Dendrogram for a five-element space (a, b, c, d, e).
Source: The author.

Hierarchical clustering techniques are classified into two categories: agglomeration-based and division-based (Fig. 4). Agglomerative algorithms, or bottom-up approach, start the analysis with as many groups as there are elements in the data space. From these initial units, groups are formed in ascending order, until all treated cases are within a single set at the end of the process. With an opposite approach, division-based algorithms, also called top-down, begin with a set that encompasses all observations, and from this initial cluster, smaller and smaller groups are formed through successive divisions. At the end of the process, there are as many groupings as cases have been treated.

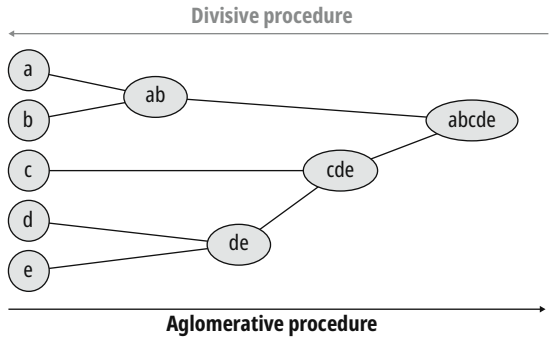

Fig. 4. Process types in hierarchical clustering.
Source: The author.

The operation scheme of agglomerative hierarchical algorithms, a classification mechanism used in our approach, is simple (Table 2). However, for its execution, it is necessary to define previously: 1) what the association actions are that will allow measuring the proximity of individuals (distance/similarity), and 2) how it can be determined when two clusters or classes can be grouped [11].

Table 2. Agglomerative hierarchical algorithm

1: Calculate the *distance matrix*
2: Define each element as a class
3: **Repeat**
4: Group the two closest classes
5: Update *distance matrix*
6: Get a single cluster as a result

Source: The author.

There are different metrics to determine the proximity of the individuals to be classified, considering their qualities. For example, if the characteristics of individuals are quantitative, a measure of distance will be used as an indicator of proximity. On the contrary, if the attributes of individuals are qualitative, a similarity index will be used as a proximity metric. The most used distances include Euclidean distance, Manhattan distance, Minkowski distance, Pearson correlation, cosine similarity, among others. The distance used in this approach is the Euclidean distance (Equation 1).

$$D^2(x_i, x_j) = (x_{1i} - x_{1j})^2 + (x_{2i} - x_{2j})^2 + \ldots + (x_{ki} - x_{kj})^2 \quad (1)$$

Having defined the proximity measure (Euclidean distance), it is necessary to define the criteria for identifying which are the closest classes to their corresponding grouping. In agglomerative hierarchical clustering, different mechanisms are distinguished to achieve this objective. These include Minimum or single-linkage clustering, Maximum or complete-linkage clustering, Mean or average-linkage clustering, and Centroid linkage clustering. Fig. 5 shows the clustering criteria used in these techniques. For example, in single-linkage, clusters are joined considering the shortest distance between the closest members of different groups, while, in complete-linkage, clusters come together considering the shortest distance between the more distant members of different groups. In the average-linkage technique, clusters are united considering the shortest average distance among all the pairs of elements of both sets [11].

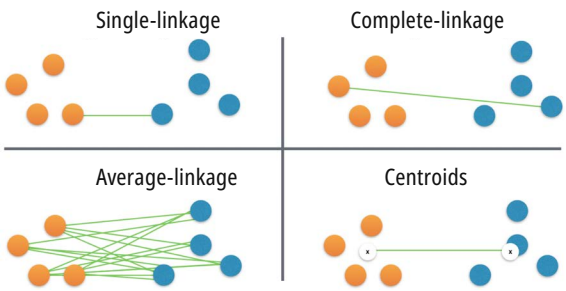

Fig. 5. Criteria for grouping in hierarchical clustering.
Source: The author.

Method

To illustrate the effectiveness of hierarchical algorithms as a mechanism for the classification of operation scenarios in power substation communications, a test scenario was designed and implemented in an isolated and controlled environment. A description of the implemented testbed, defined operating scenarios, descriptors used and the classification process carried out are discussed below.

Data capture

To capture network traffic, a prototype of a test communications network (testbed) was implemented in an isolated and controlled environment (Fig. 6). This network topology was composed of a generic interconnection device (Ethernet switch);

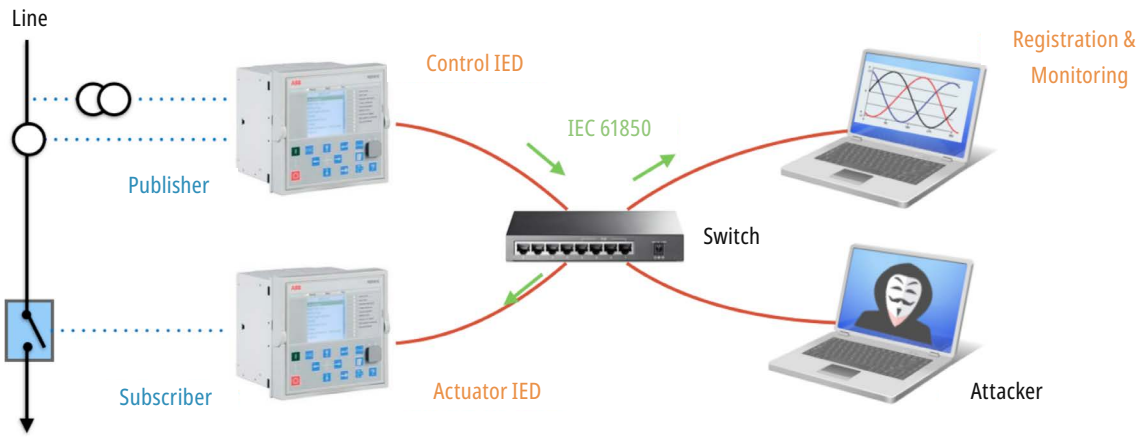

Fig. 6. Test communications network (testbed).
Source: The author.

two IEDs (ABB REM630 and ABB REG620) operating in the modes described in Fig. 6; a PC for registering and monitoring the events transmitted by IEC 61850 and the capture of network traffic through the WIRESHARK application; and a PC (attacker) to execute intrusions such as network topology discovery through the NMAP, carrying out a DoS attack using HPING3, and fabricating an spoofing attack for Goose messages sent by the publisher using the OSTINATO application. The behavior of the high voltage line was emulated by ISA DRTS66 test equipment.

On the communications network described in Fig. 6, six different operating scenarios were defined: 1) normal traffic, 2) IED disconnection, 3) network discovery attack, 4) DoS attack, 5) IED spoofing attack, and 6) failure of the high voltage line. The operating conditions of the described scenarios were generated sequentially and captured by the equipment called Registration and Monitoring in order to capture 20 minutes of traffic with 226,000 frames (PCAP file).

Descriptor identification

At this stage, a specific set of characteristics or attributes should be defined, so that each element of the data space is represented by a collection of descriptors. These descriptors will allow identifying the features that affect the classification problem. This initial choice reflects the researcher's opinion about the purpose of their classification [11].

In the field of network traffic pattern detection through unsupervised classification mechanisms, studies use two types of attributes as illustrated in Table 3: 1) attributes based on network flow, i.e., on the value of a communication protocol field (IP, UDP, TCP) or MAC address, and 2) statistical attributes, such as the average of a particular type of packet, the distribution function that the parameters of a packet perform, among others.

Our study identified that four descriptors are enough for an adequate classification of the proposed operation scenarios since we get identical results using different classification algorithms.

Table 3. Descriptors used in the recognition of traffic patterns through non-supervised algorithms

Title	Algorithm	Descriptors
Learning rules and clusters for anomaly detection in network traffic [10]	Outliers detection with k-NN	Probabilistic. P (W|U) where U = {SrcIp=128.1.2.3, DestIp=128.4.5.6} and W = {DestPort=80}
Traffic anomaly detection using k-means clustering [12]	K-means	Protocol type, source IP address, destination IP address, source port, destination port
P2P traffic identification and optimization using fuzzy c-means clustering [13]	Fuzzy c-means	NumberofPacketsSent/ReceivedforaFlow, Protocol, DurationoftheFlow, SourcePort, DestinationPort TotalNumberofPackets, MeanPacketLength, MeanPayloadLlength, MeanPacketInter-arrivalTime, AverageSent/ReceivedPacketSize, Variances, ByteRatio's
CoCoSpot: Clustering and Recognizing Botnet Command and Control Channels using Traffic Analysis [14]	Hierarchical clustering	Transport layer protocol l4p (TCP or UDP), source IP address sip, destination IP address dip, port destination dp
PeerShark: flow-clustering and conversation generation for malicious peer-to-peer traffic identification [15]	X-means (K-means that does not require previously knowing the number of classes)	Src. IP, Dest. IP, Src. port, Dest. port, Proto (TCP or UDP), Protocol, Packets per second (f/w), Packets per second (b/w), Avg. Payload size (f/w), and Avg. Payload size (b/w), with 'f/w' and 'b/w' signifying forward and backward direction of the flow, respectively
An unsupervised approach for traffic trace sanitization based on the entropy spaces [16]	K-means	StartTime, EndTime, source IP address, source port number, destination IP address, destination port number, number of packets and number of bytes in the flow.
Classification of Network Traffic Using Fuzzy Clustering for Network Security [17]	Fuzzy c-means (FCM)	Duration (seconds of the connection), src_bytes (data from src to dst), num_failed_logins, root_shell, num_access_files (operationsonaccess-controlfiles), serror_rate (percentofconnectionswith"SYN" errors), same_srv_rate (percent of connections to same service), srv_count (connections to same service in past 2 seconds).

Continued

Title	Algorithm	Descriptors
Network Intrusion Detection with Threat Agent Profiling [18]	K-means, PAM (Partitioning Around Medoids), and CLARA (Clustering LARge Application)	ID, source IP address, target IP address, category, category count, protocol, protocol count, port, duration, start timestamp, end timestamp, and ISP.
Bot detection using unsupervised machine learning [19]	K-means, X-means, and EM (Expectation-Maximization) grouping	Dstport (destinationport), maxbpktl (largestpacketsentinthebackwarddirection), maxfptkl (largest packet sent in the forward direction), fpsh cnt (times the PSH flag was set in packets traveling in the forward direction (0 for UDP)), and min fptkl (smallest packet sent in the forward direction).

Source: The author.

Table 4. Descriptors used in this study

Descriptor	Identified situation
n_frames, the average of the total number of frames captured in the time window (10 seconds)	DoS attack. This attack, independent of the service to attack, generates a huge amount of traffic on the network in a truly short period of time.
n_goose, an average of GOOSE packets captured in the time window (10 seconds)	IED Publisher disconnection or failure in the high voltage line. When an IED is disconnected, the average of GOOSE packets in the time window goes to zero. Similarly, when there is a fault in the high voltage line, the average of GOOSE messages increases as a consequence of the event.
n_arp, an average of ARP packets captured in the time window (10 seconds)	Execution of a network discovery by an intruder. Most network discovery attacks use the ARP protocol operation scheme as a strategy to discover the stations connected to the network.
goose_seqnum, SeqNum field of the GOOSE packet header	Spoofing attack of an IED Publisher. Evidence of this attack is the anomalous change of the SeqNum field values in the GOOSE header. These values are registered in sequence; therefore, any value out of order implies an intrusion. This descriptor will take the value of one if there is an anomalous change in this field. Otherwise, its value will be zero.

Source: The author.

Table 4 shows the descriptors used, three of the statistical type (n_frames, n_goose, n_arp) and one based on network flow (goose_seqnum).

Data pre-processing

Once the descriptors that will characterize the elements of our data space were identified, the traffic capture file (PCAP) was processed in order to obtain this set of elements. For this purpose, we developed a script in the LUA programming language [20] to be executed into the TSHARK application. In this way, it was possible to get a set of 110 elements. Each element, with four descriptors, shows the behavior of the network traffic in a time window of 10 seconds (Fig. 7).

Exploratory data analysis

Fig. 8 shows the behavior of normalized descriptors along the data space obtained, for each of the described operation scenarios.

Classification process

At this stage, the clustering algorithm is responsible for assigning to each element of the data space a category or class (set of elements that share certain characteristics, which also allow differentiating them from the rest). The classification process of this study was carried out using the hclust function of the software for R statistical analysis. Although there is no single criterion to determine which

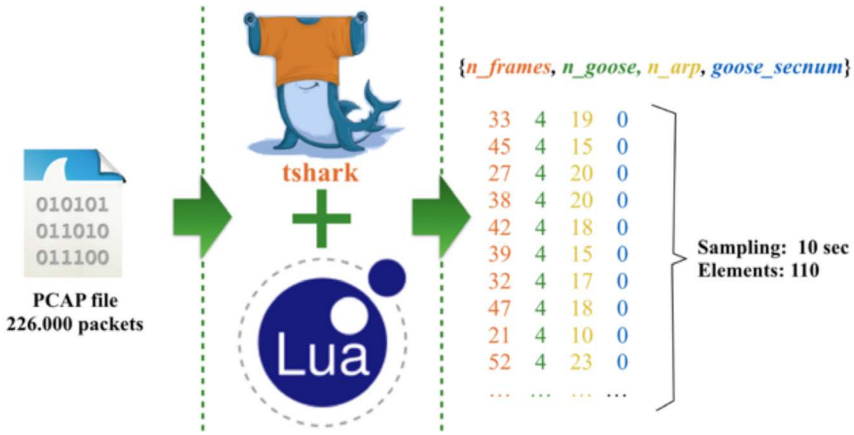

Fig. 7. Preprocessing data scheme.
Source: The author.

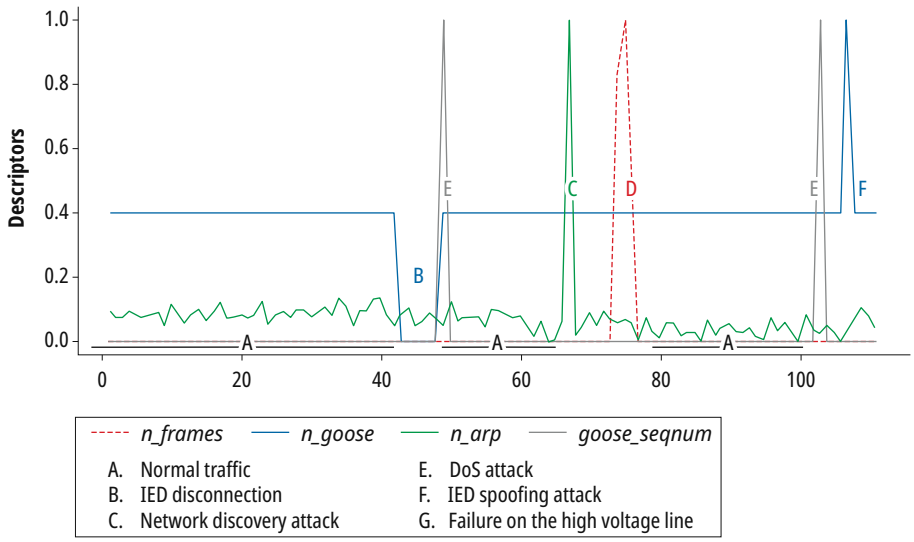

Fig. 8. Behavior of normalized descriptors.
Source: The author.

measure of association is the most appropriate to measure the proximity of individuals (distances/similarity), and which is the most convenient mechanism for grouping classes, it is recommended that results be tested and compared with different methods. Here, we opted to experiment with Euclidean distance and Gower distance as proximity metrics [21], along with single-linkage, complete-linkage, and average-linkage techniques as strategies for grouping classes. From the tests carried out, the best classification scheme was achieved using single-linkage and Euclidean distance. This combination allowed identifying the six operation scenarios described through six classes, while the other schemes required at least seven classes to correctly identify the six scenarios. Fig. 9 and Fig. 10 illustrate the structure of the dendrogram and the classification process obtained according to the behavior of descriptors.

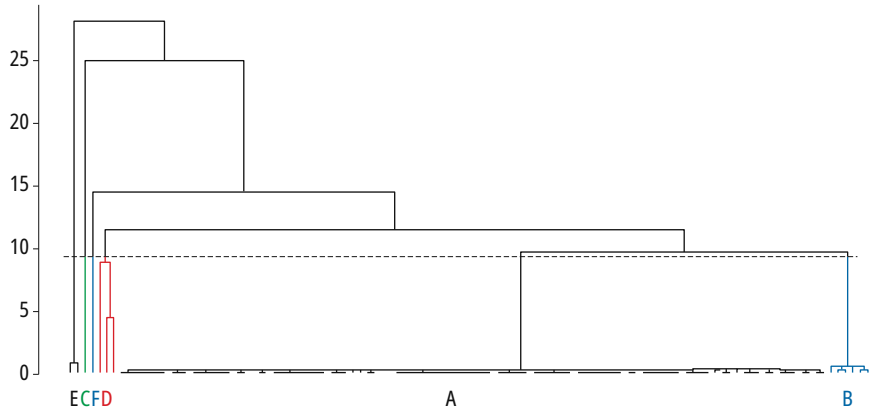

Fig. 9. Dendrogram with a six-class cut.
Source: The author.

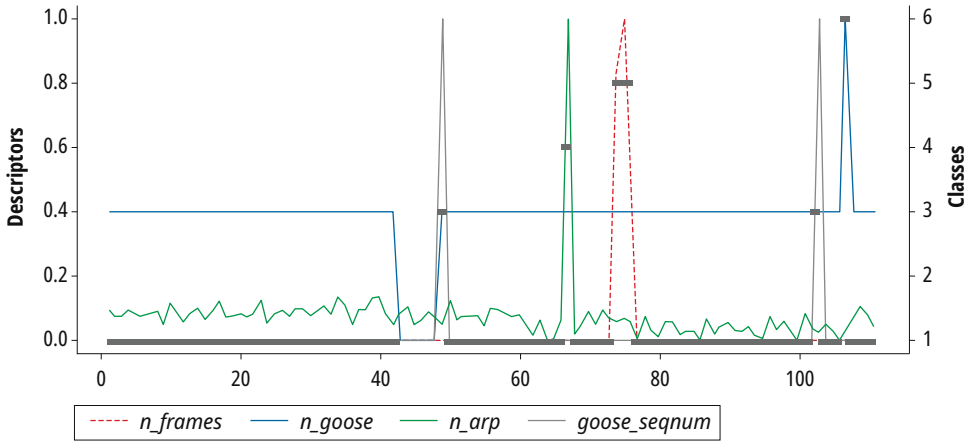

Fig. 10. Assignment of classes with the agglomerative hierarchical algorithm.
Source: The author.

Analysis of results

The tests carried out show that Gower distance and Euclidean distance exhibited similar base structures, but different clustering structures. In the same way, despite the fact that the class clustering scheme was changed (simple, complete and average), at the base of the dendrogram, it was always possible to identify each of the proposed scenarios; what changes in the structure is the way they clustered. It becomes clear that by fixing a cut-off point of the dendrogram to six or seven classes, it was possible to identify all the defined operating scenarios. However, the best result was reached using Euclidean distance with single-linkage clustering.

The analysis of the dendrogram structure (Fig. 9) shows that all operating scenarios can be clearly recognized. Yet, we expected that all the clusters related to anomalous operation scenarios were grouped in a dominant single class of failure (failure root class), i.e., completely separating the normal traffic class (A) from failure classes (B, C, D, E, F).

Validation

The results reached through an agglomerative hierarchical algorithm motivate a very qualitative interpretation, which can be subjective from the researcher's perspective. Hence, it is

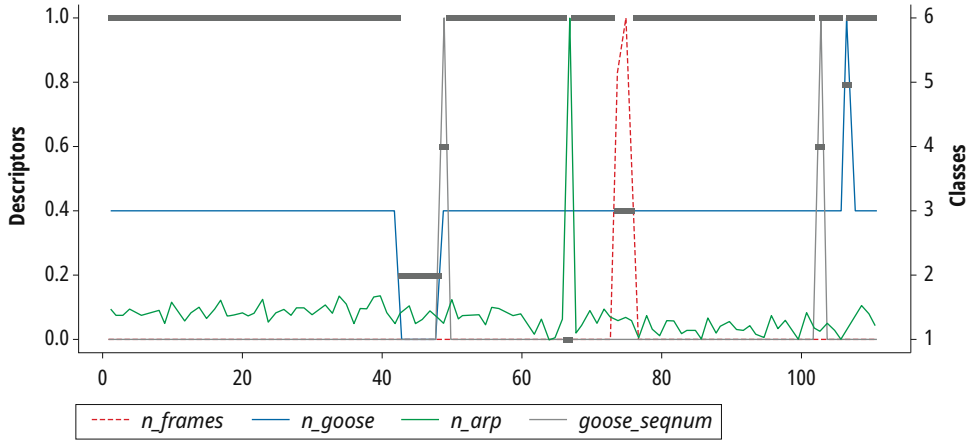

Fig. 11. Assignment of classes with the K-means partitional algorithm.
Source: The author.

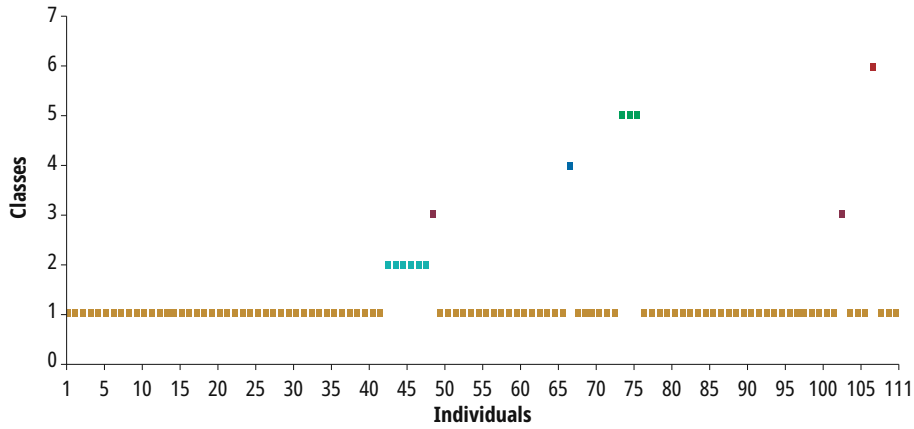

Fig. 12. Assignment of classes with the diffuse LAMDA algorithm.
Source: The author.

recommended that the achieved results be compared with other types of solutions, in which comparable results will indicate the presence of a structure in data. Thus, we explored other solution strategies using partitional and diffuse unsupervised algorithms: K-means [22] and LAMDA (Learning Algorithm Multivariable and Data Analysis) [23]. The results obtained using K-means (Fig. 11), defining in advance that the k parameter equals six (number of operation scenarios), shows how this algorithm identifies all the proposed scenarios.

Fig. 12 illustrates the classification achieved using the LAMDA algorithm with a Gaussian adaptation function, Min-Max fuzzy logic connectors and a required level of 0.6. The LAMDA algorithm is incorporated in the P3S application (DISCO Group, LAAS-CNRS), which also allows extracting the membership graphic associated with each of the classes, Global Adequacy Degree (GAD) (Fig. 13). The results obtained were in line with the K-means classification.

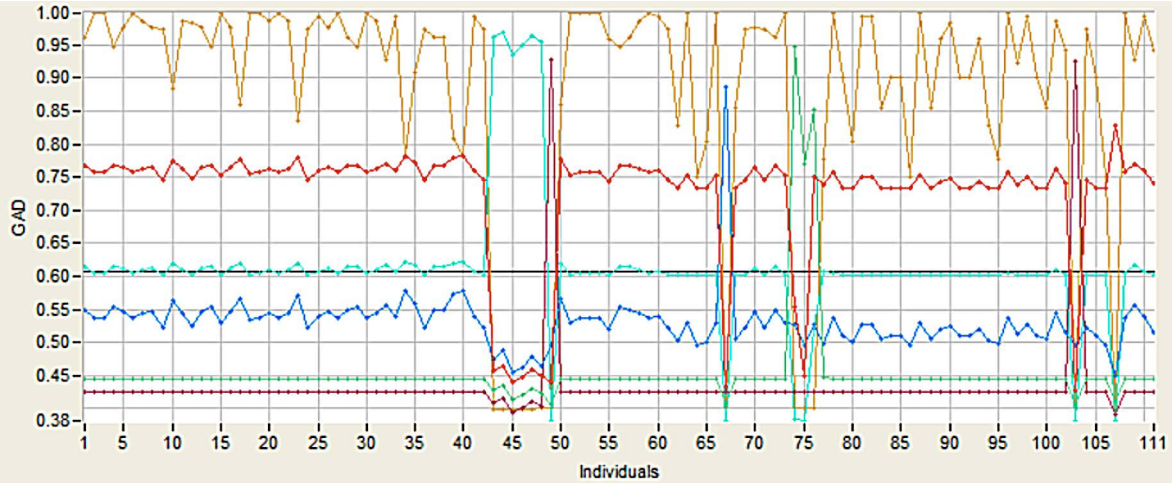

Fig. 13. Membership graph for classes in the data space.
Source: The author.

Discussion

The selection of descriptors is key in a classification process; therefore, a preliminary analysis to determine their level of contribution is necessary. Likewise, data transformation or pre-treatment may be required, just like in this work, in which three statistical descriptors (averages) and one network flow descriptor (SeqNum field in the GOOSE packet header) were used.

Through the dendrogram, the hierarchical clustering strategy allows for a preliminary exploration of possible grouping structures present in the data space, when the number of descriptors is greater than three. In this way, hierarchical clustering techniques are an excellent tool to deal with completely unknown data.

Results evidence the strength of unsupervised classification mechanisms to identify all the proposed operating scenarios by using different techniques (partitional, hierarchical and diffuse). Also, Results demonstrate that these algorithms can be useful in several scenarios, for example, traffic classification in power substation communication networks.

Finally, the fact of getting identical results from different classification algorithms proves the strength of the selected descriptors for the identification of patterns in this particular application.

Conclusion

This paper presents a practical case of how unsupervised clustering algorithms can be used as an effective tool for the identification of operation scenarios in IEC 61850 power substations communication networks. However, there are still numerous application fields to explore in this area. Particularly, the detection of new anomalies, or unknown operation scenarios, is a difficult task for a classification algorithm. In our approach, descriptor selection was successful, given the prior knowledge of operating scenarios. Descriptors proved to be robust in obtaining identical results with other unsupervised clustering techniques such as K-means (partitional-type clustering) or LAMDA (diffuse-type clustering). The challenge then is to ensure that the clustering algorithm is able to classify the normal traffic scenario in a robust manner. In this way, other scenarios will be used to notify anomalous processes in the communications network.

Aknowledgments

This work has been partially funded by the Colciencias Doctoral Fellowship (Call No. 647). I thank the Universidad de Antioquia for its financial support through the *Sostenibilidad* program. Also, I gratefully acknowledge the financial

support provided by the Colombia Scientific Program within the framework of the call *Ecosistema Científico* (Contract No. FP44842-218-2018).

References

[1] H. Farhangi, "The path of the smart grid," *IEEE power and energy magazine*, vol. 8, no. 1, pp. 18-28, 2009. https://doi.org/10.1109/MPE.2009.934876

[2] R. H. Khan and J. Y. Khan, "A comprehensive review of the application characteristics and traffic requirements of a smart grid communications network," *Computer Networks*, vol. 57, no. 3, pp. 825-845, 2013. https://doi.org/10.1016/j.comnet.2012.11.002

[3] TC57, I. E. C. "IEC 61850: Communication networks and systems for power utility automation," *International Electrotechnical Commission Std*, vol. 53, p. 54, 2010.

[4] M. T. A. Rashid, S. Yussof, Y. Yusoff, and R. Ismail, "A review of security attacks on IEC 61850 substation automation system network," in IEEE Proceedings of the 6th International Conference on Information Technology and Multimedia, Nov.2014, pp. 5-10. https://doi.org/10.1109/ICIMU.2014.7066594

[5] K. Choi, X. Chen, S. Li, M. Kim, K. Chae, and J, Na, "Intrusion detection of NSM based DoS attacks using data mining in smart grid". *Energies*, vol. 5, no. 10, pp. 4091-4109, 2012. https://doi.org/10.3390/en5104091

[6] U. K. Premaratne, J. Samarabandu, T. S. Sidhu, R. Beresh, and J. C. Tan, "An intrusion detection system for IEC 61850 automated substations." *IEEE Transactions on Power Delivery*, vol. 25, no. 4, pp. 2376-2383, 2010. https://doi.org/10.1109/TPWRD.2010.2050076

[7] J. Hoyos, M. Dehus, and T. X. Brown, "Exploiting the GOOSE protocol: A practical attack on cyber-infrastructure," in IEEE Globecom Workshops, 2012, pp. 1508-1513. https://doi.org/10.1109/GLOCOMW.2012.6477809

[8] J. Hong, C.-C. Liu, and M. Govindarasu, "Detection of cyber intrusions using network-based multicast messages for substation automation," in Innovative Smart Grid Technologies Conference (ISGT), pp. 1-5, 2014.

[9] N. Kush, E. Ahmed, M. Branagan, and E. Foo, "Poisoned goose: exploiting the goose protocol," in Proceedings of the Twelfth Australasian Information Security Conference, 2014, pp. 17-22.

[10] P. K. Chan, M. V. Mahoney, and M. H. Arshad, "Learning rules and clusters for anomaly detection in network traffic," in Managing Cyber Threats, 2005, pp. 81-99. https://doi.org/10.1007/0-387-24230-9_3

[11] J. A. Gallardo, *Análisis de datos multivariantes*. [Online]. Available: http://www.ugr.es/~gallardo/. [Accessed July 31, 2019].

[12] G. Münz, S. Li, and G. Carle, "Traffic anomaly detection using k-means clustering," in GI/ITG Workshop MMBnet, 2007, pp. 13-14.

[13] D. Liu and C. H. Lung, "P2p traffic identification and optimization using fuzzy c-means clustering," in IEEE International Conference on Fuzzy Systems (FUZZ), 2011, pp. 2245-2252. https://doi.org/10.1109/FUZZY.2011.6007613

[14] C. J. Dietrich, C. Rossow, and N. Pohlmann, "Cocospot: Clustering and recognizing botnet command and control channels using traffic analysis," *Computer Networks*, vol. 57, no. 2, pp. 475-486, 2013. https://doi.org/10.1016/j.comnet.2012.06.019

[15] P. Narang, C. Hota, and V. Venkatakrishnan, "Peershark: flow-clustering and conversation-generation for malicious peer-to-peer traffic identification," *EURASIP Journal on Information security*, vol. 2014, no. 1, p. 15, 2014. https://doi.org/10.1186/s13635-014-0015-3

[16] P. Velarde-Alvarado, C. Vargas-Rosales, R. Martinez-Pelaez, H. Toral-Cruz, and A. F. Martinez-Herrera, "An unsupervised approach for traffic trace sanitization based on the entropy spaces," *Telecommunication Systems*, vol. 61, no. 3, pp. 609-626, 2016. https://doi.org/10.1007/s11235-015-0017-6

[17] T. P. Fries, "Classification of network traffic using fuzzy clustering for network security," in Industrial Conference on Data Mining, 2017, pp. 278-285. https://doi.org/10.1007/978-3-319-62701-4_22

[18] T. Bajtoš, A. Gajdoš, L. Kleinová, K. Luˇcivjanská, and P. Sokol, "Network intrusion detection with threat agent profiling," in Security and Communication Networks, 2018.https://doi.org/10.1155/2018/3614093

[19] W. Wu, J. Alvarez, C. Liu, and H. M. Sun, "Bot detection using unsupervised machine learning," Microsystem Technologies, vol. 24, no. 1, pp. 209-217, 2018. https://doi.org/10.1007/s00542-016-3237-0

[20] R. Ierusalimschy, L. H. De Figueiredo, and W. C. Filho, "Lua-an extensible extension language," Software: Practice and Experience, vol. 26, no. 6, pp. 635-652, 1996. https://doi.org/10.1002/(SICI)1097-024X(199606)26:6<635::AID-SPE26>3.0.CO;2-P

[21] J. C. Gower, "Some distance properties of latent root and vector methods used in multivariate analysis,"

Biometrika, vol. 53, no. 3-4, pp. 325-338, 1966.https://doi.org/10.1093/biomet/53.3-4.325

[22] H. Steinhaus, "Sur la division des corp materiels en parties," Bull. Acad. Polon. Sci, vol. 1, no. 804, p. 801, 1956.

[23] J. Aguilar-Martin and R. L. De Mantaras, "The process of classification and learning the meaning of linguistic descriptors of concepts," Approximate reasoning in decision analysis, vol. 1982, pp. 165-175, 1982.