Universidad Michoacana de San Nicolás de Hidalgo

División de Estudios de Posgrado de la Facultad de Ingeniería Eléctrica.

IDENTIFICACIÓN Y CLASIFICACIÓN DE FALLAS EN LÍNEAS ELÉCTRICAS DE POTENCIA UTILIZANDO k-MEDIAS

TESIS

Que para obtener el grado de

MAESTRO EN CIENCIAS EN INGENIERÍA ELÉCTRICA

Opción Sistemas Computacionales

Presenta

Jose Guadalupe Coria Acosta

Director de Tesis

Jaime Cerda Jacobo

Doctor en Ingeniería Eléctrica y Electrónica

Morelia, Michoacán. Mayo 2023

Dedicatoria

A mi familia por haberme apoyado en todo momento y de cualquier manera que lo necesitaba. A mis maestros quienes me apoyaron en el camino del aprendizaje, brindando todos sus conocimientos que necesitaba para desarrollar el presente trabajo.

IDENTIFICACIÓN Y CLASIFICACIÓN DE FALLAS EN LINEAS ELÉCTRICAS DE POTENCIA UTILIZANDO K-MEDIAS

Los Miembros del Jurado de Examen de Grado aprueban la **Tesis de Maestría en Ciencias en Ingeniería Eléctrica** de **José Guadalupe Coria Acosta.**

Dr. José Antonio Camerena Ibarrola
Presidente del Jurado

Dr. Jaime Cerda Jacobo
Director de Tesis

Dr. Claudio Rubén Fuerte Esquivel
Vocal

Dr. Felix Calderón Solorio
Vocal

Dr. Juan Carlos Olivares Rojas
Revisor Externo (ITM)

Dr. J. Aurelio Medina Rios
Jefe de la División de Estudios de Posgrado de la Facultad de Ingeniería Eléctrica. UMSNH (Por reconocimiento de firmas)

UNIVERSIDAD MICHOACANA DE SAN NICOLÁS DE HIDALGO
Febrero 2023

Resumen

En CFE Distribución existen una gran cantidad de archivos COMTRADE(COMmon format for TRAnsient Data Exchange for power systems) que contienen la dinámica de las fallas ocurridas en el sistema de distribución eléctrica, las cuales no se encuentran etiquetados. La presente tesis propone una metodología para etiquetar y clasificar un conjunto de eventos de fallas ocurridas en líneas de transmisión eléctrica. Para esto, se utilizan algunos métodos de preprocesamiento de datos, el algoritmo de agrupamiento k-Medias y algoritmos de aprendizaje de máquina para la clasificación.

La propuesta para el etiquetamiento consiste en procesar un conjunto de archivos de formato COMTRADE que contienen mediciones del voltaje y corriente registrados durante un evento de falla e integrar las mediciones de corriente como un conjunto de datos, utilizando técnicas para mejorar la calidad de los datos como es la caracterización, la selección y la normalización de los datos, evaluadas con iVAT(Improved Visual Assessment for Tendency). Posteriormente, los conjuntos resultantes han sido agrupados en 7 grupos, usando el algoritmo de k-Medias, con la finalidad de identificar semejanzas en el conjunto de datos que permita realizar un etiquetado que esté asociado con un tipo de falla monofásica, bifásica o trifásica. Lo cual se ha logrado haciendo una comparación, mediante una matriz de confusión, con un etiquetado empírico de las fallas. Luego, para determinar las etiquetas de los tipos de fallas bifásicas aterrizadas y trifásicas aterrizadas, se ha usado el valor de la componente de secuencia cero.

Para validar la calidad del agrupamiento logrado por k-Means, se aplican varios algoritmos de clasificación, los cuales fueron entrenados con el conjunto de entrenamiento de cada etapa. Una vez entrenados, se procedió a evaluar la precisión con el conjunto de prueba correspondiente a cada una de las etapas. Los resultados de la clasificación, en cuanto a la precisión, van desde 0.38 para los datos originales del archivo COMTRADE hasta 0.9 para la versión final del pos-procesamiento. Cabe resaltar que de los clasificadores el que mejor resultado obtuvo, en el conjunto de datos pre-procecesado final, fue el SVM-RBF con 0.9, siendo ada-Boost el que peor precisión obtuvo con 0.7.

Palabras clave−*COMTRADE, Fallas en Lineas de Potencia, Preprocesamiento de datos, iVAT, k-Means*

Abstract

In CFE Distribución, there are a large number of COMTRADE (Common format for TRAnsient Data Exchange for power systems) files that contain the dynamics of faults that have occurred in the electric distribution system, which are not labeled.

This thesis proposes a methodology for labeling and classifying a set of fault events that occurred in electric transmission lines. To do this, some data preprocessing methods, the k-Means clustering algorithm, and machine learning algorithms for classification are used.

The proposed labeling method involves processing a set of COMTRADE format files containing voltage and current measurements recorded during a fault event and integrating the current measurements as a data set, using techniques to improve data quality such as characterization, selection, and normalization of data, evaluated with iVAT (Improved Visual Assessment for Tendency).

Subsequently, the resulting sets have been grouped into 7 clusters using the k-Means algorithm in order to identify similarities in the data set that allow for labeling associated with a type of single-phase, two-phase or three-phase fault. This was achieved by comparing, using a confusion matrix, with an empirical labeling of the faults. Then, to determine the labels of the grounded two-phase and three-phase faults, the value of the zero-sequence component was used.

To validate the quality of the clustering achieved by k-Means, several classification algorithms are applied, which were trained with the training set of each stage. Once trained, the accuracy was evaluated with the corresponding test set for each stage. The classification results, in terms of accuracy, range from 0.38 for the original COMTRADE file data to 0.9 for the final post-processing version. It should be noted that of the classifiers, the one that obtained the best result in the final preprocessed data set was SVM-RBF with 0.9, while ada-Boost obtained the worst precision with 0.7.

Contenido

Lista de Figuras

Lista de Tablas

Lista de Símbolos

$\mathbf{a}$	Constante vectorial
A	Amperes
A	Etiqueta para la falla monofásica de la fase a
AB	Etiqueta para la falla bifásica entre fase a y la fase b
ABC	Etiqueta para la falla trifásica
$ABCT$	Etiqueta para la falla trifásica a tierra
ABT	Etiqueta para la falla bifásica entre la fase a, la fase b y tierra
AC	Etiqueta para la falla bifásica entre fase a y la fase c
ACT	Etiqueta para la falla bifásica entre la fase a, la fase c y tierra
AF_1	Conjunto de archivos con 32 mediciones de corriente por fase
B	Etiqueta para la falla monofásica de la fase b
BC	Etiqueta para la falla bifásica entre fase b y la fase c
BCT	Etiqueta para la falla bifásica entre fase c, la fase b y tierra
$\mathbf{c}$	Centroide
C	Etiqueta para la falla monofásica de la fase c
$\mathbf{D}^*$	Matriz resultante de aplicar el algoritmo iVAT a una matriz $\mathbf{D}$
f_m	Frecuencia de Nyquist
$\mathcal{F}$	Conjunto de etiquetas del análisis empírico para 7 tipos de fallas.
FALLA	Mediciones tomadas antes de la falla.
$\mathcal{G}$	Conjunto de etiquetas del análisis empírico para 11 tipos de fallas.
$\mathbf{I}_a$	Fasor de corriente de la fase a
$\mathbf{I}_{a_0}$	Componente simétrica de secuencia cero de corriente de la fase a
$\mathbf{I}_{a_1}$	Componente simétrica de secuencia positiva de corriente de la fase a
$\mathbf{I}_{a_2}$	Componente simétrica de secuencia negativa de corriente de la fase a
$\mathbf{I}_b$	Fasor de corriente de la fase b
$\mathbf{I}_c$	Fasor de corriente de la fase c
kV	kilo Voltios
NT	Eventos que no activaron el sistema de protección
POS-FALLA	Mediciones tomadas después de la falla
PRE-FALLA	Mediciones tomadas antes de la falla
$Q_1(I_a)$	Cuartil 1 de la señal de corriente de la fase a
$Q_2(I_a)$	Cuartil 2 de la señal de corriente de la fase a
$Q_3(I_a)$	Cuartil 3 de la señal de corriente de la fase a

T	Periodo de una señal
$TRIP$	Eventos que activaron el sistema de protección
V	Voltio
$\mathbf{V}_a$	Fasor de voltaje de la fase a
$\mathbf{V}_{a_0}$	Componente simétrica de secuencia cero del voltaje de la fase a
$\mathbf{V}_{a_1}$	Componente simétrica de secuencia positiva del voltaje de la fase a
$\mathbf{V}_{a_2}$	Componente simétrica de secuencia negativa del voltaje de la fase a
$\mathbf{V}_b$	Fasor de voltaje de la fase b
$\mathbf{V}_c$	Fasor de voltaje de la fase c
$\mathbf{x}$	Vector
$\mathcal{X}$	Conjunto de conjuntos de datos
X_1^E	Conjunto de entrenamiento del conjunto X_1
X_1^P	Conjunto de pruebas del conjunto X_1
$\mathcal{X}^E$	Conjunto de conjuntos de entrenamiento
$\mathcal{X}^P$	Conjunto de conjuntos de pruebas
X_1	Conjunto de datos de la misma dimensión
X_2	Conjunto de datos X_1 reconstruido
X_3	Conjunto de datos X_2 normalizado
X_4	Conjunto de datos X_2 caracterizado y normalizado
$x(t)$	Señal analógica
$x[m]$	Seañl discreta
$\mathbf{Z}$	Impendancia
$\mathbf{Z}_f$	Impendancia de falla
$\mathbf{Z}_{ab}$	Impendancia entre la fase a y la fase b
γ	Grupo
$\Gamma^{(0)}$	Agrupamiento inicial
μ_x	Media
σ	Desviación estándar
σ^2	Varianza

Lista de Acrónimos y Abreviaturas

CFE	Comisión Federal de Electricidad
COMTRADE	COMmon format for TRAnsient Data Exchange for power systems
EHV	(Extra High Voltage) - Extra Alto Voltaje
HV	(High Voltage) - Alto Voltaje
IEEE	(Institute of Electrical and Electronics Engineers) - Instituto de ingenieros eléctricos y electrónicos
iVAT	(Improved Visual Assessment for Tendency)-Evaluación Visual de la Tendencia Mejorada
ML	(Machine Learning) - Aprendizaje de máquina
RBF	(Radial Basis Function) - Función Básica Radial
RMS	(Root Medium Square)-Raíz Cuadrática Media
SEN	Sistema Eléctrico Nacional
SEP	Sistema Eléctrico de Potencia
SME	Sistema de Monitoreo Eléctrico
SVM	(Support Vector Machine) - Máquina de Soporte Vectorial
VAT	(Visual assessment of tendency) - Evaluacíon Visual de Tendencia

Capítulo 1

Introducción

El consumo de energía eléctrica en el mundo aumenta cada vez más debido a la cantidad de dispositivos eléctricos y electrónicos que se usan para llevar a cabo las actividades cotidianas. En México, este consumo es abastecido por un sistema de potencia eléctrico trifásico que genera, transporta y distribuye la energía eléctrica a los consumidores finales. Para transportar la energía eléctrica desde el punto de generación hasta el punto de consumo se usan, básicamente, conductores eléctricos de alta tensión, aisladores y soportes. Este conjunto de elementos conforman las líneas de transmisión eléctrica.

Cuando existe una interferencia en el flujo normal de la corriente debido a un mal funcionamiento de una línea de transmisión, se dice que la línea de transmisión presenta una falla. Dicha falla puede ocasionar daños severos, tanto a las líneas de transmisión como a los dispositivos conectados en la red eléctrica.

Para evitar daños graves en la red, derivado de una falla, se instalan sistemas de protección que monitorean y, en caso de requerirse, pueden desconectar la línea de transmisión. Cada vez que los sistemas de protección detectan una falla en la red eléctrica, personal especializado analiza las mediciones tomadas durante el momento de la falla, para determinar el tipo de falla y, de ser necesario, las acciones a seguir para su reconexión. El análisis de un evento de falla suele ser tardado, e indiferentemente de sí la falla ha sido atendida o no, la información de la misma queda guardada en datos históricos.

Lo anterior conlleva a la necesidad de implementar herramientas computacionales

para reducir el tiempo de procesamiento de grandes volúmenes de datos. Generalmente, estas herramientas van encaminadas a la identificación y clasificación automática de las fallas, que en este caso están basadas en ML(Machine Learning).

En este capítulo se discuten trabajos previos encaminados a la identificación de fallas en líneas de transmisión eléctrica, con especial atención a aquellas técnicas implementadas usando ML. También se definen los objetivos y la justificación de la presente tesis y finaliza con una breve descripción del contenido de los siguientes capítulos.

1.1. Antecedentes

La detección oportuna de fallas en las líneas de transmisión eléctrica y su clasificación permite una restauración rápida del servicio y, por lo tanto, garantiza la continuidad del sistema. Es por eso que la mejora continua de las técnicas de detección de fallas y clasificación es de suma importancia.

Las técnicas de aprendizaje automático (ML) son enfoques prometedores para la clasificación de fallas, y han mostrado resultados precisos al detectar y clasificar fallas en líneas de transmisión [Tîrnovan19]. Se han utilizado técnicas como el uso de coeficientes wavelet para la detección de fallas eléctricas en tiempo real en líneas de transmisión, lo que mejora la detección en presencia de transitorios sobreamortiguados [Costa14].

Además, se ha utilizado el análisis de componentes simétricas y la transformada wavelet multinivel para clasificar con máquinas de soporte vectorial las señales de voltaje y corriente tomadas durante los eventos de falla [Jiang11]. También se ha utilizado el algoritmo de k-NN para la clasificación de fallas, el cual se ha combinado con correlación cruzada para mejorar su precisión [A. Asadi Majd17] [Aritra Dasgupta15].

Otro enfoque utilizado para la clasificación de fallas es el basado en reglas difusas separadas, el cual se utiliza para la clasificación de fallas aterrizadas y no aterrizadas mediante mediciones de corriente trifásica [R.N. Mahanty07]. Por último, se ha utilizado la transformada wavelet continua para transformar las series de tiempo de corriente y voltaje en imágenes, las cuales se utilizan para entrenar una red neuronal convolución y clasificar las fallas utilizando un árbol de decisión [Xi23].

Otra estrategia diferente ha sido implementar un algoritmo semisupervisado basado en la información conjunta de dos clasificadores supervisados, capaz de manejar datos etiquetados y no etiquetados, el cual ha mostrado ser útil para construir y entrenar clasificadores a partir de datos no etiquetados [Abdelgayed18]. Además, de manera no supervisada se ha utilizado k-Medias en conjunto con coeficientes wavelet de las señales de corriente y su componente simétrica cero para detectar y clasificar fallas, mostrando tener una muy buena precisión [Gangwar20].

En general, la clasificación de fallas utilizando técnicas de ML es ampliamente implementada, y las técnicas más utilizadas son las que pertenecen a modelos de aprendizaje supervisado, los cuales es común entrenarlos utilizando extracción de características de las señales de voltaje y corriente. Sin embargo, algunas técnicas menos exploradas, como el aprendizaje semisupervisado y no supervisado, han demostrado tener buenos resultados en la clasificación de fallas. Por lo tanto, es importante seguir investigando y mejorando estas técnicas para garantizar una detección y clasificación de fallas más precisas en las líneas de transmisión eléctrica.

1.2. Planteamiento del Problema

Actualmente, en CFE Distribución existen una gran cantidad de archivos COMTRADE(COMmon format for TRAnsient Data Exchange for power systems, por sus siglas en inglés) que contienen la dinámica de las fallas ocurridas en el sistema de distribución eléctrica, las cuales no se encuentran etiquetados. Esta falta de etiquetado impide una clasificación adecuada de las fallas presentadas en el sistema, lo que puede llevar a problemas en la detección y resolución de futuras fallas.

Para abordar este problema, es necesario desarrollar un sistema de clasificación automático que permita identificar el tipo de falla a partir de los archivos COMTRADE sin la necesidad de un análisis manual tedioso. Este sistema puede estar basado en técnicas de aprendizaje supervisado, como la clasificación mediante árboles de decisión o máquinas de soporte vectorial, o en técnicas de aprendizaje no supervisado, como la agrupación mediante el algoritmo k-Medias.

Para implementar el sistema, se requerirá un conjunto de datos etiquetados, es decir, archivos COMTRADE con las fallas clasificadas según su tipo. Este conjunto de datos puede ser generado mediante un análisis empírico inicial de una muestra representativa de los archivos, o utilizando técnicas de aprendizaje semisupervisado o no supervisado para etiquetar los archivos.

Una vez que se tenga el conjunto de datos etiquetados, se pueden entrenar modelos de aprendizaje automático para clasificar automáticamente los archivos COMTRADE. Estos modelos pueden ser integrados en un sistema que permita a los técnicos de CFE Distribución analizar y monitorear las fallas del sistema eléctrico de manera más eficiente y efectiva.

1.3. Justificación

Un sistema clasificador de fallas permitiría tomar mejores decisiones sobre las acciones a seguir después de haber ocurrido una falla, lo que contribuiría en el ahorro de gastos operativos derivados de una falla en las líneas de transmisión eléctrica. Sin embargo, esto implica que se tenga la disponibilidad de fallas etiquetadas con su tipo, lo cual actualmente no existe. El etiquetamiento preciso de un conjunto de fallas sienta las bases para la implementación de clasificadores precisos que eventualmente pudiesen ser utilizados en tiempo quasi-real.

El etiquetamiento no es preciso si se basa expresamente sobre los datos contenidos en el archivo COMTRADE. Si se toman los datos crudos de los archivos COMTRADE y se aplican algoritmos de agrupamiento, esto lleva a precisiones en los clasificadores muy bajas. Por lo anterior, se hace necesario el procesar los archivos COMTRADE y, basados en estos archivos preprocesados, extraer características que permitan un mejor agrupamiento, interpretando esto como aquel que induce una clasificación precisa.

Por otro lado, el etiquetamiento en sí, podrá ser utilizado para fines de análisis estadístico que ayuden en la toma de decisiones en cuanto a la ocurrencia de las fallas y acciones preventivas que pudieran derivarse de este análisis.

1.4. Objetivos de la tesis

1.4.1. Objetivo general

Implementar un sistema basado en ML para la identificación y etiquetado de eventos de fallas en líneas de transmisión eléctrica por su tipo de falla que haga posible la clasificación de las mismas.

1.4.2. Objetivos particulares

- Clasificar visualmente el conjunto de fallas en los registros de archivos COMTRADE.
- Procesar los archivos COMTRADE para un agrupamiento preciso.
- Etiquetar las fallas previamente preprocesadas usando k-Medias.
- Comparar el agrupamiento visual inicial con el agrupamiento generado por los algoritmos de agrupamiento.
- Evaluar la precisión del agrupamiento usando diversos clasificadores.

1.5. Metodología

La metodología implementada en esta propuesta consta de varias fases, las cuales se describen a continuación.

1.5.1. Etiquetamiento visual de las fallas

Basados en las técnicas visuales utilizadas por los encargados de atender los eventos de falla, se etiquetarán cada una de ellas en una de las once posibles:

1. Monofásica fase a a tierra (A)
2. Monofásica fase b a tierra (B)
3. Monofásica fase c a tierra (C)
4. Bifásica fase a a fase b (AB)

5. Bifásica fase a a fase c (AC)
6. Bifásica fase b a fase c (BC)
7. Bifásica fase a a fase b a tierra (ABT)
8. Bifásica fase a a fase c a tierra (ACT)
9. Bifásica fase b a fase c a tierra (ABT)
10. Trifásica (ABC)
11. Trifásica a tierra ($ABCT$)

1.5.2. Preprocesamiento de los archivos COMTRADE

Si se trata de llevar a cabo el agrupamiento de fallas utilizando los datos sin procesar almacenados en los archivos COMTRADE, el resultado puede ser poco preciso debido a la presencia de mediciones con ruido, incompletas, con diferentes escalas, entre otros problemas. En esta etapa resulta fundamental la aplicación de técnicas de preprocesamiento de datos, como la detección de valores perdidos, la normalización de datos y la reducción de dimensionalidad, ya que permiten generar agrupamientos que inducen a clasificadores más precisos y confiables.

1.5.3. Agrupamiento

Una vez que hemos preprocesado las fallas, se aplicará k-Medias como técnica de agrupamiento que induzca clasificadores precisos.

1.5.4. Validación del modelo

En esta fase, se evaluará el agrupamiento obtenido con algunos clasificadores para comprobar la precisión del agrupamiento con respecto a la clasificación. Los clasificadores implementados son: kNN, SVM Lineal, SVM RBF, Árbol de decisión, Bosque aleatorio y AdaBoost.

1.6. Descripción de los capítulos

En el capítulo 2 se realiza una revisión de los fundamentos necesarios, tanto en el área de sistemas eléctricos como en el área de ML para la implementación del presente trabajo. En el capítulo 3 se realiza una revisión de los conceptos de procesamiento de datos tomado de mediciones eléctricas y una descripción del preprocesamiento usado para del análisis de los datos contenidos en los archivos COMTRADE. En el capítulo 4 se realiza el agrupamiento y etiquetado de un conjunto de datos. En el capítulo 5 se desarrolla la implementación de los clasificadores. Por último, en el capítulo 6 se darán las conclusiones a las que se llegó durante el presente trabajo y se mencionará el alcance de trabajos futuros como resultado del trabajo actual.

Capítulo 2

Fundamentos teóricos y contextuales

En el presente capítulo se realiza una breve descripción de los fundamentos sobre los cuales esta propuesta está desarrollada. Se realizará una descripción general de los sistemas eléctricos de potencia, dando mayor énfasis en la etapa de transmisión de potencia. Específicamente se analizarán las posibles fallas que pueden ocurrir. Así mismo, estaremos evaluando su calidad en cuanto al agrupamiento, por lo cual en tal sección explicaremos algoritmo de k-Medias utilizado para el agrupamiento, así como técnicas para la evaluación de los modelos generados. En la parte final se expone trata el tema de ingeniería de características, la cual será usada para generar modelos más simples y eficientes.

2.1. Sistema eléctrico de potencia

Un Sistema Eléctrico de Potencia(SEP) cuenta, entre sus componentes más relevantes, con un conjunto de generadores, transformadores y líneas de transmisión que se encargan de abastecer la demanda eléctrica de una zona determinada[Stevenson85]. El SEP está formado principalmente por 3 etapas que son: la generación, la transmisión y la distribución. En la Figura 2.1 se muestra un esquema de un SEP general donde se observan estas 3 etapas.

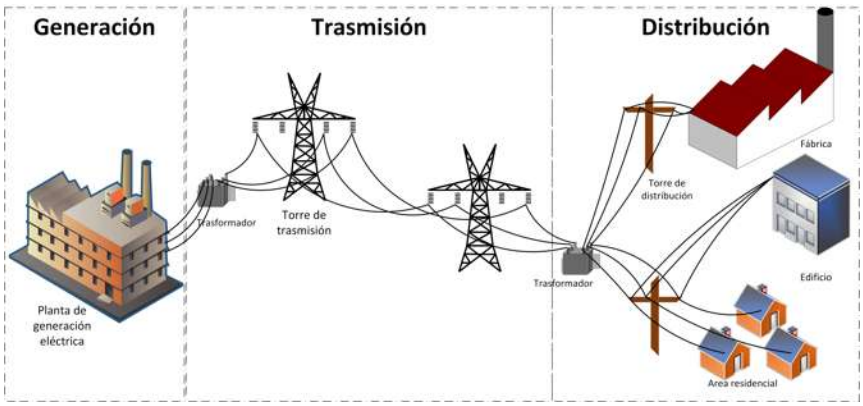
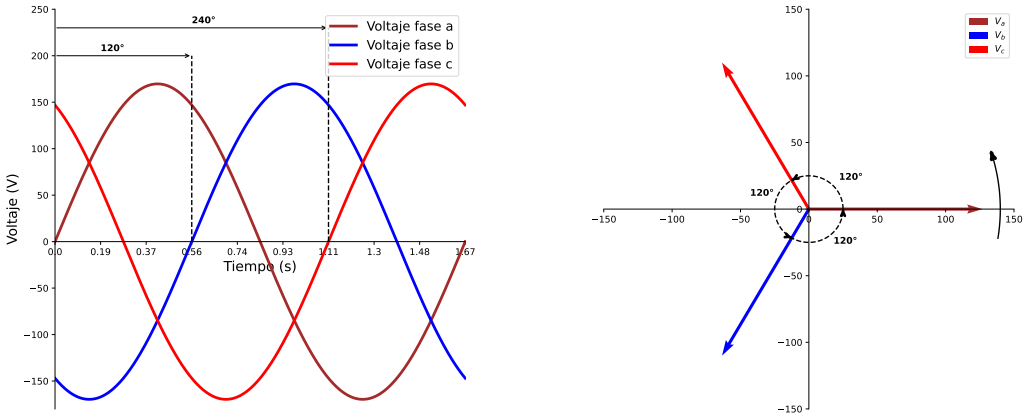


Figura 2.1: Sistema eléctrico de potencia.

El SEP que provee de energía a la República Mexicana se conoce como Sistema Eléctrico Nacional(SEN) Mexicano y, de acuerdo a la norma oficial mexicana PROY-NOM-018-CRE-2020, debe de estar diseñado para operar en una frecuencia nominal de 60Hz. En general, el SEP está basado en un sistema trifásico balanceado caracterizado por tener 3 fases, con tensiones cuya magnitud son iguales y defasadas, 120° eléctricos entre ellas.

Las 3 señales de voltaje del sistema trifásico se muestran en la Figura 2.2(a), en la cual se destaca el defase que hay de 120° entre la fase a y la fase b y de 240° entre la fase a y la fase c. Estas señales también pueden ser representadas como un sistema fasorial, tal como se muestran en la Figura 2.2(b), nótese como se señala el giro en sentido inverso a las manecillas del reloj, esto debido a que los fasores están girando en secuencia abc y un ángulo de 120° entre $\mathbf{V}_a$ y $\mathbf{V}_c$ y entre $\mathbf{V}_a$ y $\mathbf{V}_c$ de 240°. Estos diagramas son de gran utilidad para el análisis de sistemas trifásicos de potencia, tal como se expone en la sección 2.2.1.



(a) ***Señales de voltaje de un sistema trifásico*** **(b)** ***Sistema fasorial de secuencia abc***

Figura 2.2: Señales del sistema eléctrico trifásico y su representación fasorial

2.1.1. Generación

La etapa de generación es la encargada de producir la energía eléctrica necesaria para abastecer la demanda requerida por la carga conectada al SEP. La generación se divide en 2 grupos:

- **Generación centralizada**: es la generación de energía en grandes cantidades en centrales de diferentes tipos (e.g., centrales hidroeléctricas, térmicas o nucleares).

- **Generación distribuida**: es la generación de energía en pequeñas cantidades para el autoconsumo producida por las zonas residenciales y/o comerciales [CENACE20], basados en el aprovechamiento de fuentes naturales (e.g., sol, viento, corrientes de agua).

En México las principales tecnologías utilizadas para la generación de energía eléctrica son: termoeléctrica convencional, ciclo combinado, combustión interna, carboeléctrica, hidroeléctrica, geotermoeléctrica, eoloeléctrica, fotovoltaica, bioenergía, nucleoeléctrica

y cogeneración eficiente. En la Figura 2.3 se muestra el porcentaje de la capacidad tecnológica en México para la generación de la energía eléctrica, de acuerdo al tipo de tecnología utilizada.

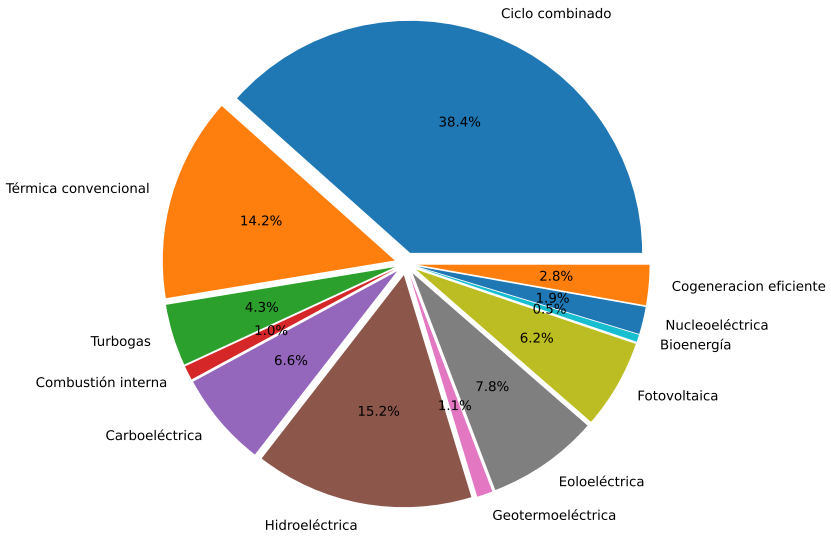


Figura 2.3: Capacidad tecnológica de generación eléctrica en México[CENACE20].

2.1.2. Transmisión

La etapa de transmisión del SEP se encarga de transportar la energía eléctrica desde su punto de producción hasta su punto de distribución para su consumo[Alvarez09], haciendo uso de líneas de transmisión aéreas y/o cables subterráneos. El sistema de transmisión transporta la energía a extra alto voltaje (EHV, por sus siglas en inglés) entre 238 kV a 400 kV y alto voltaje (HV, por sus siglas en inglés) entre 115 kV a 230 kV, con la finalidad de transmitir bajos niveles de corriente [Wildi07].

Las componentes de sistema de transmisión son:

- **Líneas de transmisión**: están compuestas por cables conductores, aisladores y estructuras de soportes. Esta infraestructura es la encargada de transportar por grandes distancias la energía eléctrica
- **Subestaciones de transmisión**: son transformadores, elevadores y reductores encargados de cambiar el valor de la magnitud de voltaje.

- **Subestaciones de interconexión**: vinculan diferentes sistemas de potencia para intercambiar potencia entre ellos.

2.1.3. Distribución

La etapa de distribución es la etapa del SEP que distribuye la energía generada a los consumidores finales de una forma segura y fiable. Esta distribución se realiza a distintos valores de tensión que van desde los 2.4kV a los 69kV para la industria y de 120V a los 600V para las zonas residenciales, comercios y pequeñas empresas. La distribución se realiza usando subestación que permiten adecuar la tensión a los valores nominales requeridos por el cliente y redes de distribución que conectan las subestaciones con el usuario final.

2.2. Fallas en líneas de transmisión

Las líneas de transmisión, tal como se mencionó en la subsección 2.1.2, forman parte de la etapa de transmisión del SEP, las cuales tienen como función transportar la energía desde un punto del sistema a otro.

Las fallas en las líneas de transmisión pueden impedir la operación de la línea eléctrica afectando al consumidor final[John J. Grainger01]. Las fallas más recurrentes son causadas por cortos circuitos. Un corto circuito ocurre cuando alguna de las fases de la línea de transmisión tienen contacto con un punto de diferente potencial eléctrico, tal como la torre de transmisión o la tierra, o cuando dos o más fases entran en contacto entre ellas, entre otros escenarios. Los cortos circuitos pueden ser provocados por deterioro en el aislamiento, descargas atmosféricas, lluvia, contaminación, corrosión, caída de árboles sobre la red eléctrica, terremotos, maniobras, etc.

Las fallas pueden provocar daños en la red eléctrica que pueden llegar a afectar los elementos eléctricos y electrónicos conectados al SEP, resultando en costosas reparaciones. Es por ello que se tienen sistemas de protección que actúan al detectar una falla en la red eléctrica. Los sistemas de protección miden continuamente los parámetros eléctricos de la red como son la corriente y voltaje. Al detectar un cambio fuera de los parámetros nominales de la red consideran que existe una falla. Si el cambio es muy abrupto, los sistemas

de protección actúan desconectando la sección de la línea donde se detectó la falla. En el caso de una desconexión, los sistemas de protección restablecen de manera automática el servicio después de algunos ciclos de operación y determinan nuevamente si aún existe la falla. Si la falla aún persiste después de la reconexión, la línea se mantiene desconectada de manera permanente hasta que un equipo técnico especializado acuda a la ubicación donde se originó la falla para realizar las actividades necesarias que restablezcan la continuidad del suministro eléctrico.

En general, el análisis de fallas lo podemos dividir en 3 partes: [Tîrnovan19]:

- **Detección**: consiste en determinar el momento de que ocurre una falla, esta parte debe ser capaz de distinguir entre un sistema en falla y uno sin falla.
- **Clasificación**: es el proceso por el cual se determina el tipo de falla ocurrido en la SEP.
- **Localización**: proceso por el cual se determina la distancia entre el punto de referencia y la ubicación de la falla.

Para el interés de esta proyecto de tesis nos enfocaremos en lo relacionado con el estudio de la detección y más a fondo en la clasificación de fallas.

2.2.1. Componentes simétricas aplicadas a fallas eléctricas

Un sistema eléctrico trifásico lo podemos representar por un sistema fasorial de 3 fasores, donde cada fasor representa una fase [Irwin97]. Si consideramos un sistema trifásico de secuencia positiva de voltajes *abc*, estos fasores se observan en la Figura 2.4

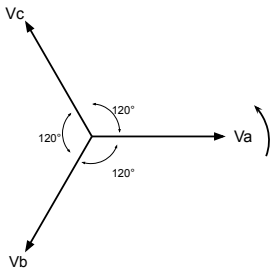


Figura 2.4: Sistema vectorial de un sistema trifásico de secuencia *abc*

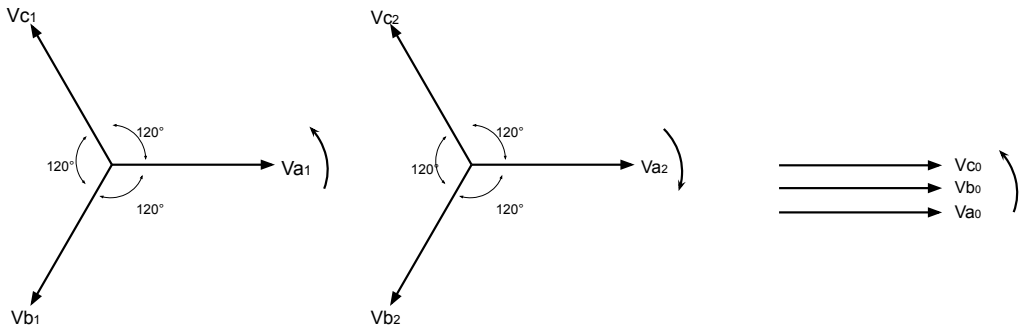
Un sistema trifásico desbalanceado se puede analizar haciendo una descomposición factorial de los fasores originales basado en el teorema de investigador C. L. Fortescue, el cual presentó ante la "American Institute of Electric Engineers", en su trabajo seminal para el estudio de sistemas polifásicos eléctricos titulado *Method of Symmetrical Coordinates Applied to the Solution of Polyphasic Networks* [Fortescue18]. En este trabajo, Fortescue demostró que un sistema desequilibrado de n vectores relacionados entre sí, pueden descomponerse en n sistemas de vectores equilibrados denominadas componentes simétricas de los vectores originales. Los n vectores de cada conjunto de componentes son de igual magnitud, siendo también iguales los ángulos formados por vectores adyacentes[Stevenson85].

Aunque este teorema es aplicable para cualquier sistema polifásico, para el caso particular de un sistema trifásico como el de la Figura 2.4, de acuerdo al Teorema de Fortecue se puede descomponer en 3 sistemas equilibrados de fasores. Los conjuntos equilibrados de fasores son:

1. Componentes de secuencia positiva: los cuales tienen la misma magnitud, con una diferencia de fase de 120° y con una secuencia de fase igual al de los originales, tal como se observa en la Figura 2.5(a).

2. Componentes de secuencia negativa: los cuales tienen la misma magnitud, con una diferencia de fase de 120° y con una secuencia de fase opuesta al de los originales, tal como se observa en la Figura 2.5(b).

3. Componentes de secuencia cero: los cuales tienen la misma magnitud, con una diferencia de fase cero, tal como se observa en la Figura 2.5(c).



(a) ***Sistema vectorial de secuencia positiva*** **(b)** ***Sistema vectorial de secuencia negativa*** **(c)** ***Sistema vectorial de secuencia cero***

Figura 2.5: Componentes simétricas

Visto de otra manera, Fortescue propone que cualquier sistema trifásico desequilibrado puede ser descrito como la suma de tres sistemas trifásicos simétricos [Fortescue18], por lo tanto, cada fasor de voltaje y/o corriente del sistema trifásico origina es igual a la sumatoria de sus componentes simétricas. Para el caso de los fasores de voltaje, seria entonces:

$$\begin{aligned} \mathbf{V}_a &= \mathbf{V}_{a_1} + \mathbf{V}_{a_2} + \mathbf{V}_{a_0} \\ \mathbf{V}_b &= \mathbf{V}_{b_1} + \mathbf{V}_{b_2} + \mathbf{V}_{b_0} \\ \mathbf{V}_c &= \mathbf{V}_{c_1} + \mathbf{V}_{c_2} + \mathbf{V}_{c_0} \end{aligned} \tag{2.1}$$

Para obtener el valor de cada vector de secuencia es necesario resolver el sistema de ecuaciones 2.1. Las componentes simétricas de secuencia positiva y negativa están defasados $120^{\underline{o}}$, por lo cual, si definimos $\mathbf{a} = 1\underline{/1}20°$, dicho sistema da como resultado el sistema de ecuaciones (2.2):

$$\begin{aligned} \mathbf{V}_a &= \mathbf{V}_{a_1} + \mathbf{V}_{a_2} + \mathbf{V}_{a_0} \\ \mathbf{V}_b &= \mathbf{a}^2\mathbf{V}_{a_1} + \mathbf{a}\mathbf{V}_{a_2} + \mathbf{V}_{a_0} \\ \mathbf{V}_c &= \mathbf{a}\mathbf{V}_{a_1} + \mathbf{a}^2\mathbf{V}_{a_2} + \mathbf{V}_{a_0} \end{aligned} \tag{2.2}$$

que en su forma matricial se puede expresar como:

$$\begin{bmatrix}\mathbf{V}_a\\ \mathbf{V}_b\\ \mathbf{V}_c\end{bmatrix} = \begin{bmatrix}1 & 1 & 1\\ 1 & \mathbf{a}^2 & \mathbf{a}\\ 1 & \mathbf{a} & \mathbf{a}^2\end{bmatrix}\begin{bmatrix}\mathbf{V}_{a_0}\\ \mathbf{V}_{a_1}\\ \mathbf{V}_{a_2}\end{bmatrix} \tag{2.3}$$

resolviendo para las variables del sistema de componentes simétricas obtenemos:

$$\begin{bmatrix}\mathbf{V}_{a_0}\\ \mathbf{V}_{a_1}\\ \mathbf{V}_{a_2}\end{bmatrix} = \frac{1}{3}\begin{bmatrix}1 & 1 & 1\\ 1 & \mathbf{a} & \mathbf{a}^2\\ 1 & \mathbf{a}^2 & \mathbf{a}\end{bmatrix}\begin{bmatrix}\mathbf{V}_a\\ \mathbf{V}_b\\ \mathbf{V}_c\end{bmatrix} \tag{2.4}$$

si volvemos a reescribir en su forma original, este sistema de ecuaciones es:

$$\begin{aligned}\mathbf{V}_{a_0} &= \frac{1}{3}(\mathbf{V}_a + \mathbf{V}_b + \mathbf{V}_c)\\ \mathbf{V}_{a_1} &= \frac{1}{3}(\mathbf{V}_a + \mathbf{a}\mathbf{V}_b + \mathbf{a}^2\mathbf{V}_c)\\ \mathbf{V}_{a_2} &= \frac{1}{3}(\mathbf{V}_a + \mathbf{a}^2\mathbf{V}_b + \mathbf{a}\mathbf{V}_c)\end{aligned} \tag{2.5}$$

alternativamente para la corriente las ecuaciones son:

$$\begin{aligned}\mathbf{I}_{a_0} &= \tfrac{1}{3}(\mathbf{I}_a + \mathbf{I}_b + \mathbf{I}_c)\\ \mathbf{I}_{a_1} &= \tfrac{1}{3}(\mathbf{I}_a + \mathbf{a}\mathbf{I}_b + \mathbf{a}^2\mathbf{I}_c)\\ \mathbf{I}_{a_2} &= \tfrac{1}{3}(\mathbf{I}_a + \mathbf{a}^2\mathbf{I}_b + \mathbf{a}\mathbf{I}_c)\end{aligned} \tag{2.6}$$

de manera matricial se puede reescribir como:

$$\begin{bmatrix}\mathbf{I}_{a_0}\\ \mathbf{I}_{a_1}\\ \mathbf{I}_{a_2}\end{bmatrix} = \frac{1}{3}\begin{bmatrix}1 & 1 & 1\\ 1 & \mathbf{a} & \mathbf{a}^2\\ 1 & \mathbf{a}^2 & \mathbf{a}\end{bmatrix}\begin{bmatrix}\mathbf{I}_a\\ \mathbf{I}_b\\ \mathbf{I}_c\end{bmatrix} \tag{2.7}$$

Por ejemplo, si tenemos un sistema trifásico con corrientes $\mathbf{I}_a = 30\underline{/15}°\text{A}$, $\mathbf{I}_b = 60\underline{/-150}°\text{A}$ e $\mathbf{I}_c = 45\underline{/150}°\text{A}$ y se pretende obtener las componentes simétricas de las corrientes de fase, primero se obtienen los componentes simétricas para la fase a, sustituimos los valores de las corrientes en Ecuación (2.7), dando como resultado:

$$\begin{bmatrix}\mathbf{I}_{a_0}\\ \mathbf{I}_{a_1}\\ \mathbf{I}_{a_2}\end{bmatrix} = \frac{1}{3}\begin{bmatrix}1 & 1 & 1\\ 1 & \mathbf{a} & \mathbf{a}^2\\ 1 & \mathbf{a}^2 & \mathbf{a}\end{bmatrix}\begin{bmatrix}30\underline{/15°}\\ 60\underline{/210°}\\ 45\underline{/150°}\end{bmatrix} = \begin{bmatrix}21.88\underline{/185.99°}\\ 38.53\underline{/3.32°}\\ 14.54\underline{/32.51°}\end{bmatrix}$$

Por otro lado, las componentes simétricas de secuencia, para las fases b y c, son de la misma magnitud que las de fase a. Sin embargo, los ángulos son distintos para el caso de las componentes simétricas positivas y negativas. Por lo tanto, las componentes simétricas de la fase b y c quedan como:

$$\begin{bmatrix} \mathbf{I}_{b_0} \\ \mathbf{I}_{b_1} \\ \mathbf{I}_{b_2} \end{bmatrix} = \begin{bmatrix} 21.88\angle 185.99° \\ 38.53\angle 3.32° + 240° \\ 14.54\angle 32.51° + 120° \end{bmatrix} = \begin{bmatrix} 21.88\angle 185.99° \\ 38.53\angle 243.32° \\ 14.54\angle 152.51° \end{bmatrix}$$

$$\begin{bmatrix} \mathbf{I}_{c_0} \\ \mathbf{I}_{c_1} \\ \mathbf{I}_{c_2} \end{bmatrix} = \begin{bmatrix} 21.88\angle 185.99° \\ 38.53\angle 3.32° + 120° \\ 14.54\angle 32.51° + 240° \end{bmatrix} = \begin{bmatrix} 21.88\angle 185.99° \\ 38.53\angle 123.32° \\ 14.54\angle 272.51° \end{bmatrix}$$

En la Figura 2.6 se muestra el diagrama fasorial de las componentes simétricas resultado del ejemplo anterior. Los fasores originales desbalanceados están de color verde, violeta y verde azulado, mientras que las componentes simétricas de secuencia cero de color cian, la componente de secuencia positiva en amarillo y la componente de secuencia negativa en azul marino. Se puede notar que la suma vectorial de las componentes de secuencia es el vector original.

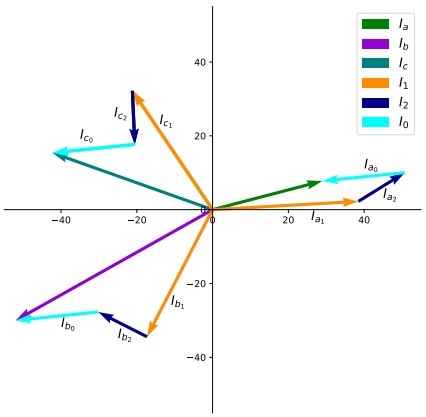


Figura 2.6: Relación geométrica entre los componentes de fase y las componentes de secuencia

2.2.2. Clasificación de fallas

La clasificación de fallas en líneas de transmisión la podemos dividir en dos grupos:

- **Fallas simétricas**: son aquellas fallas donde las tres fases entran en corto circuito o las tres fases entran en contacto con la tierra, es decir, son dos 2 tipos:

  1. Trifásica (ABC)
  2. Trifásica a tierra ($ABCT$)

- **Fallas asimétricas**: son aquellas fallas donde interviene una fase, dos fases o dos fases a tierra. Este tipo de fallas son las más comunes en el SEP y pueden ser uno de los siguientes 9 tipos:

  1. Monofásica de la fase a a tierra (A)
  2. Monofásica de la fase b a tierra (B)
  3. Monofásica de la fase c a tierra (C)
  4. Bifásica de la fase a y la fase b (AB)
  5. Bifásica de la fase a a fase c (AC)
  6. Bifásica de la fase b a fase c (BC)
  7. Bifásica de la fase a, a la fase b y a tierra (ABT)
  8. Bifásica de la fase a, a la fase c y a tierra (ACT)
  9. Bifásica de la fase b, a la fase c y a tierra (BCT)

por lo tanto, tenemos 11 tipos de fallas de las cuales son 3 son monofásicas, 3 bifásicas, 3 bifásicas a tierra, 1 trifásica y 1 trifásica a tierra.[Stevenson85]

Fallas monofásicas a tierra

Las fallas monofásicas es un tipo de falla asimétrica, la cual consiste en el contacto de una de las 3 fases del sistema trifásico con tierra o una impedancia de bajo valor. Este tipo de fallas son las más frecuentes en las líneas de transmisión. En la Figura 2.7 se ilustra

un diagrama de 3 fases *abc* donde la fase *c* se encuentra en falla puesta a tierra por una impedancia baja Z_f.

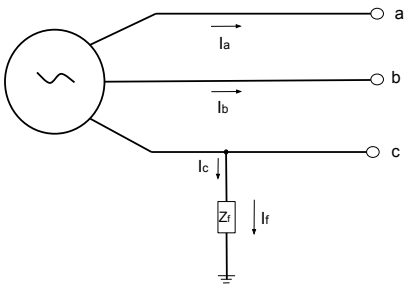


Figura 2.7: Falla monofásica a tierra

Cuando ocurre esta falla se puede asumir que $Z_f \backsimeq 0$, por lo que la corriente $\mathbf{I}_c$ es muy elevada y las corrientes $\mathbf{I}_a$ y $\mathbf{I}_b$ son despreciables debido a que son mucho menores que $\mathbf{I}_c(\mathbf{I}_c \gg \mathbf{I}_a, \mathbf{I}_c \gg \mathbf{I}_b)$. Por lo tanto, las condiciones iniciales para el análisis de una falla monofásica son:

$$\begin{aligned} \mathbf{V}_c &\backsimeq 0 \\ \mathbf{I}_a &= \mathbf{I}_b = 0 \\ \mathbf{I}_c &= \mathbf{I}_f = \frac{\mathbf{V}_c}{Z_f} \end{aligned}$$

Si sustituimos estas ecuaciones iniciales en 2.6 obtenemos:

$$\begin{aligned} \mathbf{I}_{a_0} &= \frac{1}{3}\mathbf{I_a} \\ \mathbf{I}_{a_1} &= \frac{1}{3}\mathbf{I_a} \\ \mathbf{I}_{a_2} &= \frac{1}{3}\mathbf{I_a} \end{aligned} \tag{2.8}$$

Por lo que $\mathbf{I}_{a_0} = \mathbf{I}_{a_1} = \mathbf{I}_{a_2} = \frac{1}{3}\mathbf{I}_a$

Fallas bifásicas

Las fallas bifásicas pertenecen al conjunto de fallas asimétricas, las cuales consisten en el contacto directo de 2 de las 3 fases del sistema o por una impedancia de bajo valor entre las fases. En la Figura 2.8 se ilustra un diagrama de 3 fases *abc* donde las fases *b* y *c* se encuentra en contacto a través de impedancia baja $\mathbf{Z}_{bc}$.

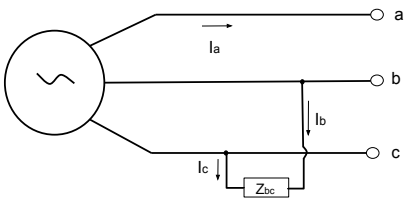


Figura 2.8: Falla bifásica

Cuando ocurre esta falla se puede asumir que $\mathbf{Z}_{ab} \backsimeq 0$ por lo que las corrientes de las fases en "contacto" b y c serán de igual magnitud, pero en sentido contrario y la $\mathbf{I}_a$ será despreciable por ser mucho menor a $\mathbf{I}_b$ e $\mathbf{I}_c (\mathbf{I}_b \gg \mathbf{I}_a \ll \mathbf{I}_c)$. Mientras que el voltaje de $\mathbf{V}_b$ y $\mathbf{V}_c$ serán iguales. Por lo tanto, las condiciones de falla son:

$$\begin{aligned} \mathbf{V}_b &= \mathbf{V}_c \\ \mathbf{I}_a &= 0 \\ \mathbf{I}_c &= -\mathbf{I}_b \end{aligned}$$

Si sustituimos estas condiciones iniciales en 2.6 obtenemos:

$$\begin{aligned} \mathbf{I}_{a_0} &= 0 \\ \mathbf{I}_{a_1} &= \frac{1}{3}(\mathbf{a}\mathbf{I}_b - \mathbf{a}^2\mathbf{I}_b) \\ \mathbf{I}_{a_2} &= -\frac{1}{3}(-\mathbf{a}^2\mathbf{I}_b + \mathbf{a}\mathbf{I}_a) \end{aligned} \tag{2.9}$$

Por lo que $\mathbf{I}_{a_0} = 0$ y $\mathbf{I}_{a_1} = -\mathbf{I}_{a_2}$

Fallas bifásicas a tierra

Las fallas bifásicas es un tipo de falla asimétrica, la cual consiste en el contacto directo de 2 líneas o por una impedancia de bajo valor y también en contacto con tierra. En la Figura 2.9 se ilustra un diagrama de 3 fases abc donde las fases b y c se encuentra en contacto a través de impedancia baja Z_f que tiene contacto con tierra.

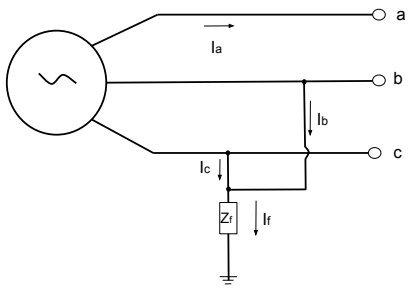


Figura 2.9: Falla bifásica a tierra

Cuando ocurre esta falla se puede asumir que $Z_f \backsimeq 0$, por lo que las corrientes de las fases en "contacto" b y c serán de igual magnitud y la $\mathbf{I}_a$ será despreciable por ser mucho menor que $\mathbf{I}_b$ e $\mathbf{I}_c$. Por lo tanto, las condiciones de falla son:

$$\mathbf{V}_b = \mathbf{V}_c = 0$$

$$\mathbf{I}_a = 0$$

Si sustituimos estas condiciones iniciales en 2.5 obtenemos:

$$\mathbf{V}_{a_0} = \frac{1}{3}\mathbf{V}_a$$
$$\mathbf{V}_{a_1} = \frac{1}{3}\mathbf{V}_a$$
$$\mathbf{V}_{a_2} = \frac{1}{3}\mathbf{V}_a$$

Es decir $\mathbf{V}_{a_0} = \mathbf{V}_{a_1} = \mathbf{V}_{a_2}$. Por otro lado, las componentes simétricas de corriente estarán dadas por:

$$\mathbf{I}_{a_0} = \frac{1}{3}(\mathbf{I}_b + \mathbf{I}_c)$$
$$\mathbf{I}_{a_1} = \frac{1}{3}(\mathbf{a}\mathbf{I}_b + \mathbf{a}^2\mathbf{I}_c)$$
$$\mathbf{I}_{a_2} = \frac{1}{3}(\mathbf{a}^2\mathbf{I}_b + \mathbf{a}\mathbf{I}_c)$$

Fallas trifásicas

Las fallas trifásicas pertenecen al grupo de las fallas simétricas y suceden cuando las 3 fases de la línea entran en contacto directo entre ellas o por medio de una impedancia

de bajo valor. El diagrama de este tipo de fallas se puede observar en la Figura 2.10, del cual se puede notar que las impedancias entre cada una de las líneas y su punto de contacto está nombrada como $\mathbf{Z}_{abc}$, lo que indica que las tres impedancias son idealmente iguales.

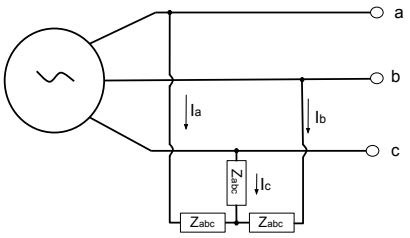


Figura 2.10: Falla trifásica

El circuito formado durante este tipo de falla se trata de un circuito eléctrico trifásico balanceado, lo que permite calcular los voltajes y corrientes analizando solo una fase[Irwin97]. Para calcular el voltaje en $\mathbf{Z}_{abc}$ usaremos la ley de Ohm[Ohm27], que aplicada a este problema en la fase a se formula como:

$$\mathbf{V}_a = \mathbf{I}_a \mathbf{Z}_{abc} \tag{2.10}$$

Dado que, idealmente la impedancia $\mathbf{Z}_{abc}$ es igual a cero, podemos calcular $\mathbf{V}_a$ sustituyendo en la Ecuación (2.10) $\mathbf{Z}_{abc} = 0$, por lo cual, $\mathbf{V}_a = 0$. Por la simetría del sistema tenemos los voltajes medidos en las otras fases en el punto de falla son:

$$\mathbf{V}_a = \mathbf{V}_b = \mathbf{V}_c = 0$$

Para obtener las componentes simétricas de voltaje para las fallas trifásicas se calculan usando la Ecuación (2.5), quedando como resultado:

$$\begin{aligned} \mathbf{V}_{a_0} &= 0 \\ \mathbf{V}_{a_1} &= 0 \\ \mathbf{V}_{a_2} &= 0 \end{aligned} \tag{2.11}$$

Para obtener las corrientes hacemos uso del siguiente sistema de ecuaciones derivadas de

un análisis de componentes simétricas usando el teorema de Thévenin[Mujal14]:

$$\begin{aligned} \mathbf{V}_{a_0} =& \mathbf{E} - \mathbf{I}_{a_1}\mathbf{Z}_1 \\ \mathbf{V}_{a_1} =& -\mathbf{I}_{a_2}\mathbf{Z}_2 \\ \mathbf{V}_{a_2} =& -\mathbf{I}_{a_0}\mathbf{Z}_0 \end{aligned} \tag{2.12}$$

donde E es una tensión equivalente de Thévenin y $\mathbf{Z}_1, \mathbf{Z}_2, \mathbf{Z}_3$ son las impedancias de secuencia positiva, negativa y cero vistas desde el punto donde se ha dado el cortocircuito.

Si sustituimos 2.11 en 2.12 y despejamos las corrientes de secuencia obtenemos que:

$$\begin{aligned} \mathbf{I}_{a_0} &= 0 \\ \mathbf{I}_{a_1} &= \frac{E}{\mathbf{Z}_1} \\ \mathbf{I}_{a_2} &= 0 \end{aligned} \tag{2.13}$$

En general, cuando existe una falla por corto circuito, el voltaje medido en el punto de corto circuito y dependerá de la corriente que fluya por la impedancia de falla. Mientras que la corriente en la fase afectada incrementara de manera súbita. Por otro lado, las componentes simétricas de secuencia cero de corriente se presentan cuando hay flujo de corriente a tierra, mientras que las corrientes de secuencia negativa se presentan cuando el sistema eléctrico se encuentre en desbalance [Mujal14].

2.3. Registro de fallas de líneas de transmisión

Los Sistemas de Monitoreo Eléctrico (SME), entre otras funciones, permiten obtener mediciones de señales de voltaje y corriente en tiempo real. Estas mediciones son tomadas a una frecuencia determinada y programada en el SME. Dicha programación puede variar de acuerdo a las características de la línea de transmisión y necesidades del usuario del sistema de monitoreo.

Cuando sucede una falla en la red eléctrica, las mediciones hechas por el SME son procesadas en tiempo real para su detección y determinación de la acción a tomar. Además, es almacenada con la finalidad de hacer un análisis posterior al evento que ayude a la toma de decisiones sobre ajustes, mantenimiento o reparación del las líneas de transmisión.

2.3.1. Almacenamiento de eventos en formato COMTRADE

Una falla eléctrica es también conocida como evento de falla y generalmente, en el contexto adecuado, se le llama solamente evento. Los datos registrados por la SME durante un evento comúnmente son almacenados en el formato estandarizado por el IEEE llamado COMTRADE(Common Format for Transient Data Exchange). Este formato consiste en los siguientes 4 archivos:

1. **Archivo de datos (.dat)**: Archivo de texto obligatorio en formato ASCII. Contiene el registro de las mediciones, así como también la hora y fecha en que fue tomada cada medición.

2. **Archivo de configuración (.cfg)**:Archivo de texto obligatorio en formato ASCII. Este archivo contiene la información para interpretar los registros de las mediciones, tal como, tasa de muestreo y número de señales.

3. **Archivo de cabecera (.hdr)**: Archivo de texto opcional en formato ASCII. Contiene información implementaría que describe el sistema eléctrico del que fueron tomadas las mediciones, como puede ser: descripción de las fallas o tensiones nominales.

4. **Archivo de información (.inf)**: Contiene información en formato que pueda ser utilizada por equipos y software del fabricante o de otros fabricantes.

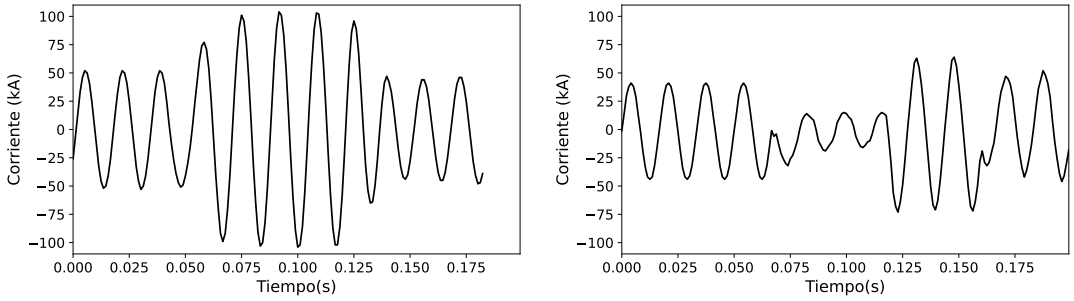
2.3.2. Datos de medición en sistemas eléctricos

Cuando los datos de los eventos son capturados por distintos SMEs, es probable que tengan métodos de almacenamiento distintos. Por lo cual, si la intención es comparar distintos eventos, es importante conocer cuáles son las diferencias entre los datos almacenados por los distintos SMEs.

Enseguida se comentará sobre algunas de las diferencias encontradas en los datos almacenados en distintos SMEs y se sugieren algunas soluciones para poder comparar estos datos aunque sean generados por distintos SMEs.

Duración del evento

El sistema de registro almacena los valores de voltaje y corriente durante un periodo de tiempo conocido como duración del evento y es programado por el SME. Como se puede apreciar en la Figura 2.11 se muestran dos señales de diferente magnitud pero con diferente duración de evento. En la Figura 2.11(a) la duración del evento es de 0.1822 s, mientras que en la Figura 2.11(b) es de 0.1989s.



(a) ***Señal con 0.1822s*** **(b)** ***Señal con 0.1989s***

Figura 2.11: Señales con diferente duración

Cuando se desea comparar punto a punto estas series de tiempo, puede resultar en un problema debido a que en algunos casos los datos serán insuficientes. Si se desea hacer esta comparación lo podemos abordar varias formas de las cuales se pueden mencionar las siguientes:

- **Eliminando datos a una cantidad propuesta.** Se eliminan los datos al inicio o al final de la señal, teniendo en cuenta aquellos que no representen una gran importancia para el análisis de la señal. Esta técnica puede hacer que se pierdan datos valiosos si no se evalúa correctamente qué datos son los importantes.

- **Predicción de datos.** Aplicar algún método de regresión a los datos para predecir los datos que continúen y poder completar series a la misma longitud. Esta técnica puede hacer que se genere alguna tendencia o estacionalidad que no necesariamente sea útil para el análisis.

- **Repetir los últimos registros.** Se puede completar usando los últimos registros de las series de tiempo, ejemplo puede ser el último ciclo. Esto permite traer las últimas características de la falla que se pueden considerar la parte estable del registro.

- **Alineación de señales**: otro enfoque es el utilizado por técnicas que permitan alinear las señales, calculando la similitud entre dos señales que varíen su velocidad, tales como Deformación Dinámica del Tiempo (DTW, por sus siglas en inglés) o su variante ADTW.

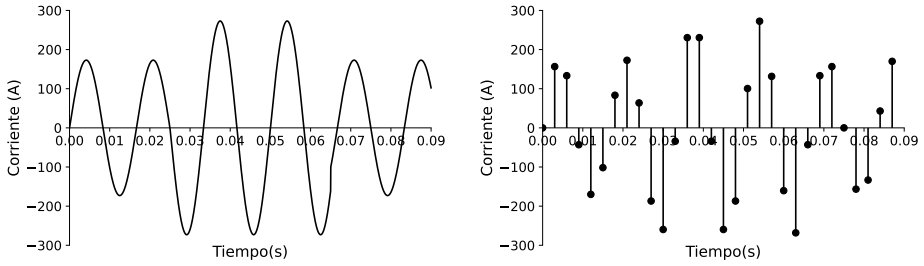
Muestreo de una señal

El muestreo es la discretización de una señal que tiene como objetivo la conversión de una señal analógica a una discreta. Esto se logra haciendo el registro del valor de magnitud de la señal $x(t)$ cada cierto periodo T, por lo cual podemos definir el muestreo de una señal matemáticamente como la multiplicación de una señal analógica por un tren de pulsos, dando como resultado una señal discreta $x[m]$, es decir:

$$x[m] = x(t)\delta(t - mT) \tag{2.14}$$

donde $\delta(t)$ es una delta de dirac.

Para ilustrar el resultado del muestreo de una señal analógica, se presentan dos gráficos en la Figura 2.12. La gráfica mostrada en la Figura 2.12(a) se trata de una señal continua que oscila a una frecuencia de 60 Hz, mientras que la otra gráfica apreciada en la Figura 2.12(b) es el resultado de muestrear la señal continua a una frecuencia de 333 Hz. Como se puede observar en la señal discreta, solo se representan algunos valores de la señal analógica, lo que puede provocar pérdida de información.



(a) *Señal continua* **(b) *Señal discreta***

Figura 2.12: Muestreo de señales

La frecuencia a la que se muestrea una señal se le conoce como tasa de muestreo. Para que una señal muestreada pueda ser reconstruida en la original de manera precisa, de acuerdo al teorema de muestreo de Nyquist, la tasa de muestreo f_s debe de ser al menos el doble de la frecuencia máxima de la señal original f_m, tal como se muestra en la Ecuación (2.15).

$$f_s \geq 2f_m \tag{2.15}$$

La tasa de muestreo puede variar y algunos motivos por los cual puede suceder son parámetros del SME, de la frecuencia de la señal a muestrear o el problema que se desee resolver.

Identificación del inicio y final de la falla

Los datos de un evento almacenado del SME pueden ser divididos en tres tipos: PRE-FALLA, FALLA y POS-FALLA. Esto se puede observar en la Figura 2.13, la cual está seccionada en tres partes. La primera de ella es antes de que suceda la falla (PRE-FALLA), donde se observan mediciones de la señal sin valores atípicos, a partir de que sucede la falla sigue un subconjunto de datos atípicos (FALLA); en este caso son altos valores de corrientes mayores al máximo de la zona PRE-FALLA. Finalmente, continua una zona conocida como POS-FALLA, que es cuando la corriente regresa a sus valores típicos que son iguales a la zona PRE-FALLA.

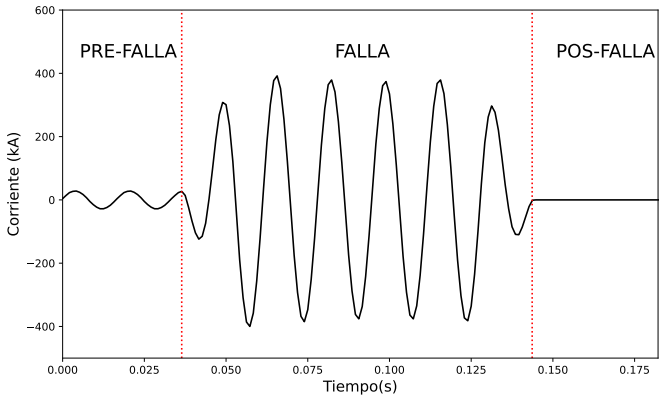


Figura 2.13: Partes de una señal de falla.

En algunas ocasiones los valores de la línea no regresan a los valores óptimos, por lo que se considera que en la línea se ha mantenido la falla. En otros casos, el almacenamiento de los datos de la falla no contiene la zona PRE-FALLA, es por ello que es necesario determinar cada una de estas secciones antes de procesar los datos almacenados, con la finalidad de hacer un análisis correcto.

Señales con disparo del sistema de protección

Existen anomalías en las señales de corriente y/o voltaje tales que se salen de los parámetros máximos de operación de las líneas de transmisión eléctrica (e.g., aumento muy grande en la magnitud de corriente, descenso de la magnitud de voltaje, cambios en la fase). Dichas fallas pueden ser tan graves que suelen dañar las líneas de transmisión eléctrica haciendo muy costosa su reparación. Existen sistemas de protección que actúan en el momento en que sucede evento, tomando decisión previamente programada para proteger la línea. Uno de ellos es la protección de sobre corriente que actúa desconectado la línea de trasmisión cuando los valores de corriente son mayores a los permitidos por este sistema de protección, en algunos casos, esta desconexión se da después de ciertos ciclos que se mantiene la falla.

Al desconectar las líneas de transmisión, el monitoreo de la señal de corriente

debería ser idealmente cero. Un evento de este tipo se puede observar en la Figura 2.14 donde se muestra en la zona de POST-FALLA con un valor de 0 A constante.

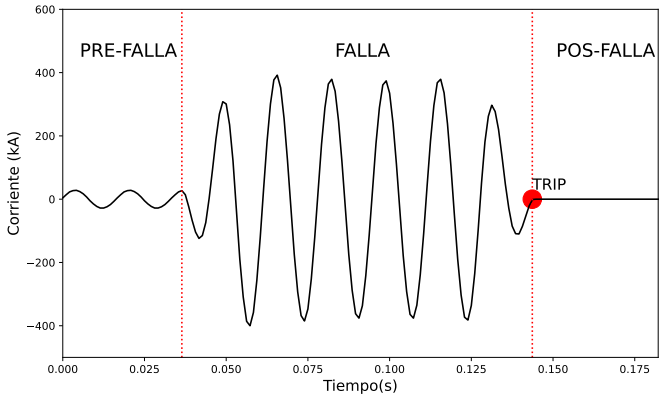


Figura 2.14: Señal con disparo en el sistema de protección.

2.4. Agrupamiento con K-Medias

El agrupamiento (Clustering) es una técnica que busca grupos de objetos semejantes que están más relacionados entre sí, que con objetos de otros grupos [Raschka15]. En análisis de datos, los objetos son los datos, los cuales pueden ser imágenes, sonidos, mediciones, respuestas, etc., también conocidos como vectores de atributos.

Las técnicas de agrupamiento en aprendizaje de máquina, buscan agrupar datos, de un conjunto de datos, que desconoce a priori la cantidad de grupos en los que se puede asociar el conjunto de datos.

k-Medias es un algoritmo de agrupamiento que divide un conjunto de datos X en k grupos. Donde $X = \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, ..., \mathbf{x}_n\}$, n es la cantidad de datos y cada elemendo $\mathbf{x} \in R^d$ (d la dimensión de los datos).

Cada grupo, al que nombraremos como γ_j, donde j es un valor entre 1 y k, está representado por un centroide $\mathbf{c}$, que es la media aritmética del grupo. Este algoritmo que se puede resumir en los siguientes pasos:

1. Se determina la cantidad de grupos en los que se dividirá el conjunto de datos X.

2. Se inicializa el valor de los centroides(C) en el espacio dimensional de los datos.

3. Se asocia cada dato a un grupo usando la distancia más cercana entre el dato y el centroide.

4. Se redefine el valor de los centroides usando la media aritmética de cada grupo.

5. Se repite el paso 3 y 4 hasta que los grupos no cambien.

En seguida se mencionan 3 de los métodos más comúnmente usados para el cálculo de la distancia entre los centroides y los datos. El símbolo que usaremos para nombrar a la distancia es L.

1. **Distancia de Minkowski**: es una medida de distancia entre dos puntos en un espacio vectorial.
$$L(\mathbf{x}, \mathbf{c}) = \left(\sum_{i=1}^{r} |\mathbf{x}_i - \mathbf{c}_i|^p \right)^{\frac{1}{p}} \tag{2.16}$$
Con $p > 1$.
En particular, la distancia de Minkowski p=1 se le conoce como la distancia Manhattan y con $p = 2$ se le conoce como la distancia Euclidiana.

2. **Distancia de correlación**. se basa en la correlación de 2 vectores la cual mide la dirección.
$$L(\mathbf{x}, \mathbf{c}) = 1 - \frac{\langle \mathbf{x} - \mathbf{c} \rangle}{||\mathbf{x}||||\mathbf{c}|} \tag{2.17}$$

3. **Distancia coseno**: se basa entre el coseno del ángulo de 2 vectores.
$$L(\mathbf{x}, \mathbf{c}) = 1 - \frac{\mathbf{x} \cdot \mathbf{c}}{||\mathbf{x}|| \cdot ||\mathbf{c}||} \tag{2.18}$$

En el Algoritmo 1 describe un pseudocódigo para la implementación de K-Medias, dando la posibilidad de hacer el cálculo con cualquiera de las distancias anteriormente mencionadas.

En la Figura 2.15 se muestra la aplicación del Algoritmo 1 sobre un conjunto de datos. El conjunto de datos están representados por puntos en el espacio de 2 dimensiones, como se muestra Figura 2.15(a). Se asume que el conjunto de datos se puede dividir en

Algoritmo 1 Algoritmo k-Medias

Input : k - numero de grupos

$X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ - Conjunto de datos de n elementos

$C = \{\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_k\}$, - Conjunto de k elementos que representa los centroides

$\Gamma^{(0)} = \{\gamma_1, \gamma_2, \ldots, \gamma_k\} = \emptyset$ - Agrupamiento en k grupos con m datos semejantes en el tiempo $t = 0$

Output: $\mathbf{\Gamma}^{(t)}$

repeat

$t = t + 1$

for $i \leftarrow 1$ *to* n **do**

$distancia_1 = L(X_i, C_1)$

$p = 1$

for $j \leftarrow 2$ *to* k **do**

$distancia_j = L(X_i, C_j)$

if $distancia_j < distancia_{j-1}$ **then**

$p = j$

end

end

$\Gamma_p^{(t)} \cup \{X_i\}$

end

for $j \leftarrow 1$ *to* k **do**

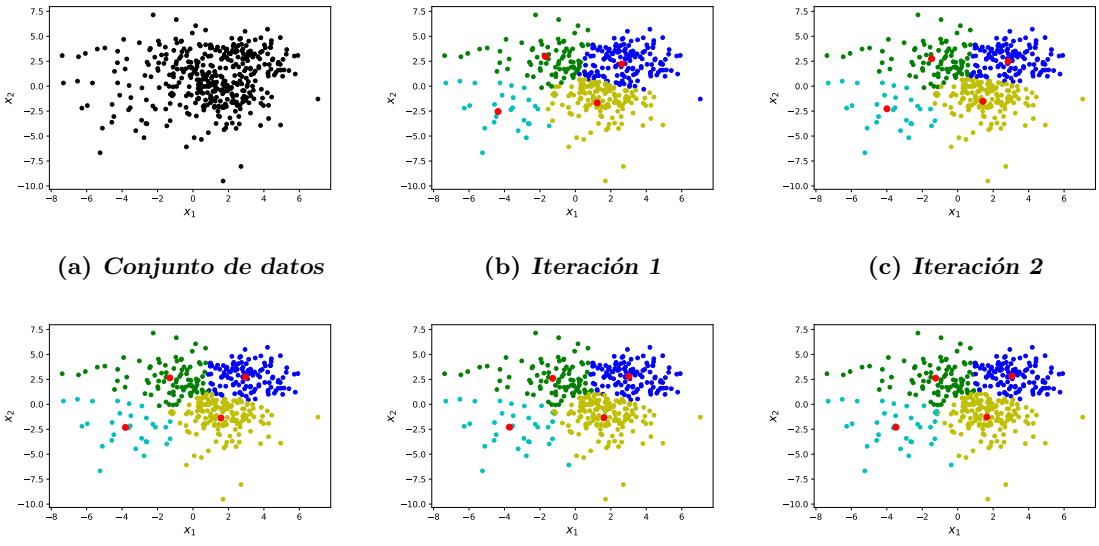
$C_j^{(t)} = \frac{\sum_{i=1}^{n} \Gamma_{ij}^{(t)}}{n}$

end

until $C^{(t)} = C^{(t-1)}$;

return $\Gamma^{(t)}$

4 grupos, por lo tanto, el valor de k tendrá el mismo valor. Al aplicar el algoritmo busca iterativamente un buen agrupamiento, tal como se observa muestra de la Figura 2.15(b) a la Figura 2.15(f), donde se observa como van cambiando el agrupamiento de figura a figura, representado los centroides por un punto de color rojo y cada grupo por un color distinto. Logrando su objetivo en 5 iteraciones.



(a) ***Conjunto de datos*** **(b)** ***Iteración 1*** **(c)** ***Iteración 2***

(d) ***Iteración 3*** **(e)** ***Iteración 4*** **(f)** ***Iteración 5***

Figura 2.15: Agrupamiento usando k-Medias

Para implementar k-Medias, de antemano, se requiere conocer o tener una idea de que en cantidad de k grupos se puede dividir el conjunto de datos. También el resultado obtenido del agrupamiento es muy sensible a la inicialización de los centroides, ya que diferente selecciones de centroides puede derivar en diferentes resultados de agrupamiento.

Aunque existen otros métodos de agrupamiento que no requieren inicialmente conocer el valor de k(i.e. el algoritmo DBSCAN(Density-Based Spatial Clustering of Applications with Noise)[Ester96] o agrupamiento jerárquico[Leonerd90]), en el presente trabajo se ha usado como método de agrupamiento únicamente el algoritmo k-Medias. Esto se debe a que se conoce a priori la cantidad de grupos (i.e. k) en los que se puede dividir la base de datos con la que se ha trabajado, derivado de la teoría descrita en la sección 2.2.2. Además,

este algoritmo es ampliamente utilizado en aplicaciones de aprendizaje de máquina por ser fácil de entender e implementar.

2.4.1. Métodos para estimar el valor de k

Uno de los principales problemas de k-Medias es estimar el valor de grupos k en los que se puede dividir el conjunto de datos. Como regla general, el valor k debe de ser menor o igual a los n números de datos con los que se cuentan. Para determinar el valor de k nos podemos valer de algunas herramientas tales como los dendrogramas, método coeficiente de la silueta, VAT e iVAT.

Dendrograma

Es una representación visual de un conjunto de datos dada por un proceso de agrupación jerárquica. Para hacer esta representación se parte de un conjunto de datos $X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ de n elementos. Inicialmente, se divide el conjunto de datos en dos grupos, tales que, cada elemento dentro de cada grupo este vinculado entre sí, después cada uno de 2 conjuntos se divide en otros 2 grupos más pequeños, y así sucesivamente hasta que queden n conjuntos. Para determinar que datos formarán parte del nuevo grupo al dividirse se considera la distancia entre los datos. Mientras que para su vinculación, hay varios métodos para su cálculo, de los cuales los más frecuentemente usados son:

1. Enlace único: es la distancia más cercana entre los datos. La construcción dendrogramas a partir de este tipo de enlace es sensible a datos atípicos.

2. Enlace completo: es la distancia más lejana entre los datos. La construcción de dendrogramas a partir de este enlace produce grupos más compactos y bien separados, además de ser menos sensible a datos atípicos.

3. Enlace de centroides: es la distancia entre centroides.

4. Enlace Warn: es la diferencia de la suma cuadrática de la distancia de cada dato al centroide del grupo menos la suma cuadrática de la distancia entre los centroides de los grupos más cercanos.

Para hacer el gráfico se considera que, cada vez que se hace una división de datos, se genera una línea horizontal y dos verticales que unen a los siguientes grupos. A modo de ejemplo, supongamos que tenemos los datos mostrados en la Figura 2.16(a), de los cuales se pretende realizar un dendrograma. Para ello es necesario determinar cuál será el método de vinculación y la métrica de distancia. Si se selecciona como método de vinculación el enlace Warn y la medida de distancia a la distancia euclidiana, el resultado obtenido es el mostrado en la Figura 2.16(b).

El método para la selección del valor de k usando un dendrograma consiste en:

1. Identificar la línea vertical de mayor longitud.

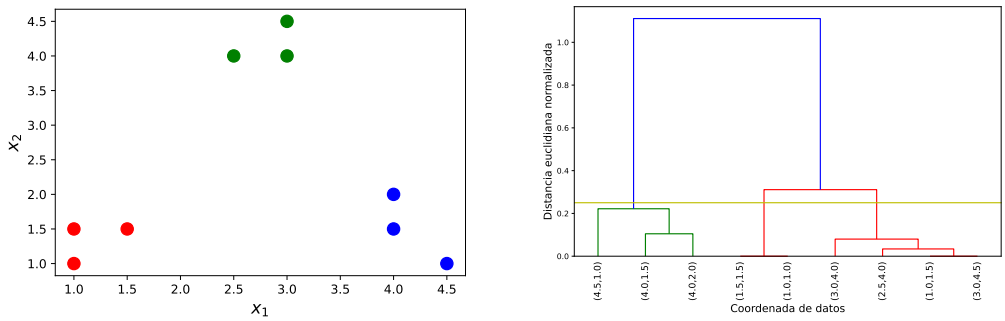
2. En donde termina la línea más larga, se traza una línea horizontal sobre todo el gráfico.

3. Se cuentan las líneas verticales que son cruzadas con la línea horizontal trazada.

4. El valor del conteo será el valor de k propuesto.

identificar donde se ubica la línea vertical de mayor longitud.

En la Figura 2.16(b) se muestra la línea horizontal trazada de color amarillo y como se puede notar, cruza 3 líneas verticales, 1 azul y 2 rojas, por lo cual, el valor de k para este ejemplo es de 3.



(a) ***Conjunto de datos*** **(b)** ***Dendrograma***

Figura 2.16: Nivel de agrupamiento usando dendrograma

Un dendrograma puede ser útil para explorar la estructura de los datos y determinar el número aproximado de grupos en el análisis de agrupamientos, puede ser menos práctico en conjuntos de datos grandes debido a su complejidad computacional. Además, si la dimensionalidad de los datos es alta, el dendrograma puede ser difícil de interpretar debido a su complejidad visual.

Método del codo

El método del codo consiste en maximizar la varianza entre los grupos y minimizar la varianza dentro de los grupos del agrupamiento. Para ello es necesario calcular la inercia del agrupamiento usando la Ecuación (2.19).

$$WCSS = \sum_{r=1}^{k} \sum_{s=1}^{|\gamma_r|} d(\gamma_{rs}, \mathbf{c}_r) \tag{2.19}$$

donde γ_r es un grupo, $|\gamma_r|$ la cantidad de elementos del grupo γ_r y $\mathbf{c}_r$ el centroide. El método consiste en calcular la inercia del agrupamiento($WCSS$) y graficarlos con distintos valores de k, que van desde 1 hasta el número total de elementos en el conjunto de datos, tal como se muestra en la Figura 2.17. En la cual observa un cambio en la dirección con un punto rojo en k igual a 3 haciendo forma de un codo, a este punto se le asocia con la cantidad de grupos que contiene el conjunto de datos [Géron19].

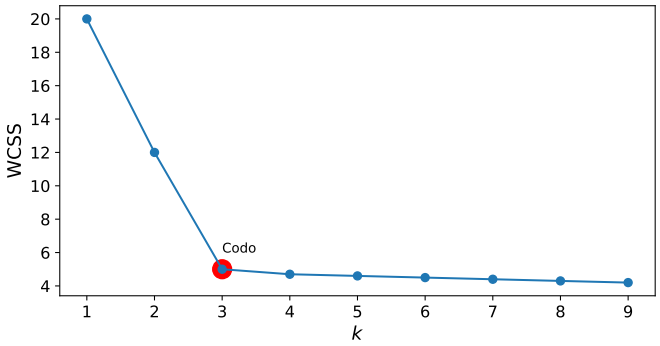


Figura 2.17: Método del codo

En algunos casos el codo en el gráfico no es muy claro, esto puede deberse a que en

un conjunto de datos haya ruido, existan muchos grupos, no se adapte el modelo o existan pocos datos.

Método de silueta

El método de silueta mide la similitud entre los datos dentro de un mismo grupo y los datos de otros grupos [Géron19]. El método consiste en calcular el coeficiente de silueta usando la siguiente ecuación:.

$$S(\mathbf{x}_i) = \frac{g(\mathbf{x}_i) - f(\mathbf{x}_i)}{max(f(\mathbf{x}_i), g(\mathbf{x}_i))} \tag{2.20}$$

donde $f(\mathbf{x}_i)$ es la distancia promedio entre un dato $\mathbf{x}_i$ y los demás datos de su mismo grupo, $g(\mathbf{x}_i)$ es la distancia promedio entre el dato $\mathbf{x}_i$ al los datos contenidos en el grupo más cercano y $max(f(\mathbf{x}_i), g(\mathbf{x}_i))$ es el máximo valor entre $f(\mathbf{x}_i)$ y $g(\mathbf{x}_i)$. Se entiende que si $S(\mathbf{x}_i)$ es igual a 1 es bueno el agrupamiento y si es -1 es malo. El coeficiente de silueta para un agrupamiento será entonces:

$$SC = \frac{1}{n} \sum_{i=1}^{n} S(\mathbf{x}_i) \tag{2.21}$$

Al igual que el método del codo se hace el cálculo de SC para diferentes valores de k, tomando como número de grupos el valor de k tal que el SC sea el más elevado. Un ejemplo es presentado en la Figura 2.18, donde se observa una gráfica de valores de SC calculados para varios valores dé k. Se nota un punto rojo en (3,4) donde se visualiza el mayor valor del coeficiente de silueta, por lo cual, para conjunto de datos analizado, el valor sugerido de k es 3.

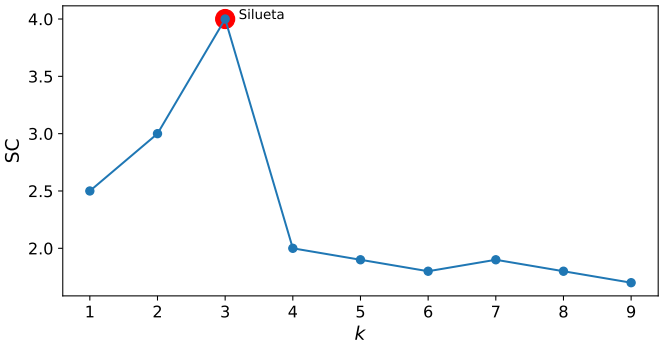


Figura 2.18: Método de la silueta.

2.4.2. Evaluación visual del agrupamiento: VAT e iVAT

Los métodos VAT (Visual assessment of tendency) e iVAT(Incremental Visual Assessment of the Clustering Tendency) permiten representar patrones en gráficos 2D para evaluar el nivel del agrupamiento de un grupo de datos. Una interpretación adecuada del gráfico resultante, ayudará en mejor el modelo de agrupamiento.

VAT e iVAT parten de la matriz de distancia, también conocida como la matriz de similitud [J.C. Bezdek22]. Cada elemento de la matriz representa la distancia que hay entre un elemento de un conjunto de datos y otro elemento dentro del mismo conjunto. Tomando como ejemplo los puntos los datos de Figura 2.16(a) tenemos una matriz de disimilitud tal como se muestra a continuación:

$$\begin{bmatrix} 0.00 & 0.50 & 0.50 & 0.50 & 3.60 & 0.00 & 3.20 & 0.50 & 3.53 \\ 0.50 & 0.00 & 0.70 & 0.70 & 3.35 & 0.50 & 2.91 & 0.00 & 3.04 \\ 0.50 & 0.70 & 0.00 & 0.00 & 4.03 & 0.50 & 3.60 & 0.70 & 3.50 \\ 0.50 & 0.70 & 0.00 & 0.00 & 4.03 & 0.50 & 3.60 & 0.70 & 3.50 \\ 3.60 & 3.35 & 4.03 & 4.03 & 0.00 & 3.60 & 0.50 & 3.35 & 3.80 \\ 0.00 & 0.50 & 0.50 & 0.50 & 3.60 & 0.00 & 3.20 & 0.50 & 3.53 \\ 3.20 & 2.91 & 3.60 & 3.60 & 0.50 & 3.20 & 0.00 & 2.91 & 3.35 \\ 0.50 & 0.00 & 0.70 & 0.70 & 3.35 & 0.50 & 2.91 & 0.00 & 3.04 \\ 3.53 & 3.04 & 3.50 & 3.50 & 3.80 & 3.53 & 3.35 & 3.04 & 0.00 \end{bmatrix}$$

Con dicha matriz podemos hacer un gráfico con mapa de colores en escala de grises tal como

se muestra en la en la Figura 2.19 donde entre más oscuro sea el color menor es valor del elemento de la matriz.

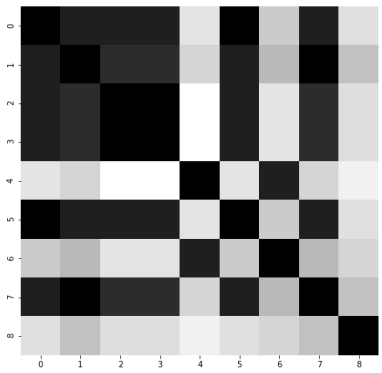
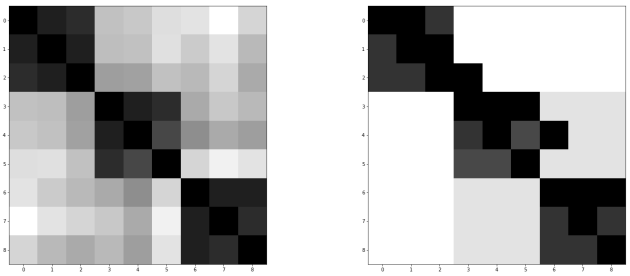


Figura 2.19: Mapa de colores de matriz de similitud.

Más concretamente, los métodos de VAT (Visual assessment of tendency) son algoritmos de evaluación de tendencia que se basan en la matriz de distancias (disimilitud) de un conjunto de datos y después generan una imagen monocromática reordenada, o un mapa de color, que muestra posibles conglomerados en los datos en bloques oscuros a lo largo de la diagonal. el algoritmo de VAT se describe en el Algoritmo 2.

Se han hecho modificaciones en el algoritmo para mejorar la visualización de los conglomerados sobre la diagonal, por lo que hay toda una familia de algoritmos VAT. Uno de ellos es algoritmo iVAT (Improved Visual Assessment for Tendency), el cual trasforma la matriz de disimilitud en una medida de distancia basada en rutas. iVAT mejorando significativamente la visualización del gráfico obtenido con VAT, este algoritmo es mostrado en el Algoritmo 3[Kumar20].

A modo de ejemplo, podemos considerar los datos en la Figura 2.16(a), si aplicamos los Algoritmos 2 y 3 obtendríamos como resultado la Figura 2.20 donde se observan los gráficos VAT e iVAT. Como se puede notar, son evidentes 3 grupos grandes sobre la diagonal. Mostrando una alta tendencia al agrupamiento este conjunto de datos.



(a) *Gráfico VAT* **(b) *Gráfico iVAT***

Figura 2.20: Gráficos VAT e iVAT.

Estos métodos de visualización son utilizados previo al agrupamiento para analizar si existe un grupo dentro de un conjunto de datos. Y posterior al agrupamiento se utilizan para evaluar que tan bien está agrupado un grupo en particular.

Algoritmo 2 Algoritmo VAT

Input : D—$n \times n$ matriz de disimilitud que satisface

—$D_{ij} \geq 0$

—$D_{ij} = D_{ji} \forall i, j$

—$D_{ii} = 0 \forall i$

Output: D^*—$n \times n$ matriz ordenada de disimilitud

$I(D^*)$—Imagen VAT de D^*

P—Índices VAT de reordenamiento de D

d—ordenación de magnitudes de corte MST

Poner $K = \{1, 2, ..., n\}$, $I = J = 0$

Seleccionar $(i, j) \in \underset{k \in K q \in K}{argmax} D_{kq}$

$P_1 = i$; $I = \{i\}$ y $J = K$—$\{i\}$

for $t \leftarrow 2$ **to** n **do**

  Seccionar $(i, j) \in \underset{k \in I q \in J}{argmin} D_{kq}$

  $P_t = j; I \cup \{j\}; J = J$—$\{j\}; d_{t-1} = D_{ij}$

end

for $p \leftarrow 1$ **to** n **do**

  for $q \leftarrow 1$ **to** n **do**

    $D^*_{p,q} = D_{P_p, P_q}$

  end

end

Crear $I(D^*)$

Algoritmo 3 Algoritmo iVAT

Input : D—$n \times n$ matriz de disimilitud reordenada VAT

Output: D'^{*}—$n \times n$ matriz de disimilitud iVAT

$I(D^{*})$—Imagen VAT de D^{*}

for $r \leftarrow 2$ **to** n **do**

$j = \underset{1 \leq k \leq r-1}{argmax}\{D^{*}_{rk}\}$

$D'^{*}_{rj} = D^{*}_{rj}$

$c = \{1, 2, ..., r-1\}$—$\{j\}$

$D'^{*}_{rc} = max\{D'^{*}_{rj}, D'^{*}_{jc}\}$

end

$D'^{*}_{rc} = D'^{*}_{cr}$

2.5. Transformaciones de datos

La mayoría de los algoritmos de agrupamiento y clasificación tienen un mejor desempeño si los datos son numéricamente comparables o están en la misma escala. Es por ello que se sugiere hacer un preprocesamiento de datos para mejorar el nivel de agrupamiento. Existen algunas técnicas para escalar y normalizar los datos. A continuación se mencionarán dos de ellos usados especialmente en clasificadores basados en medidas de distancia euclidiana [Raschka15].

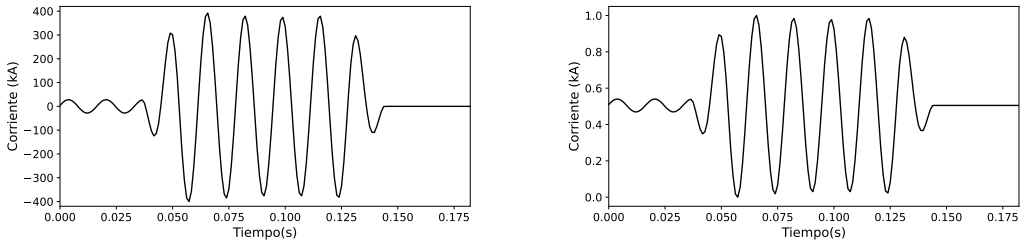
2.5.1. Escalamiento de datos

Se usa cuando se desea cambiar un conjunto de datos $\mathbf{x}$ que se se encuentra en el rango $[\bar{x} \ldots \underline{x}]$ por otro rango y en el rango $[\bar{y} \ldots \underline{y}]$ dando como resultado un nuevo conjunto de datos $\boldsymbol{y}$. La Ecuación (2.22) permite calcular el escalamiento.

$$\boldsymbol{y} = \underline{y} + \frac{(\mathbf{x} + \underline{x})\left(\bar{y} - \underline{y}\right)}{\bar{x} - \underline{x}} \tag{2.22}$$

Donde $\underline{y}$ es el valor mínimo al que se quiere escalar, $\bar{y}$ es el valor máximo al que se quiere escalar, $\underline{x}$ es valor mínimo del conjunto original y $\bar{x}$ es el valor máximo del conjunto original.

La Figura 2.21 muestra un ejemplo de este método de escalamiento. En la Figura 2.21(a) se muestra la señal original, mientras que la Figura 2.21(b) se muestra la señal escalada entre un rango de -1 a 1, cuando se usa este rango en particular se le conoce como normalización. Nótese como se mantiene la misma forma de onda en diferentes escalas.



(a) *Señal* **(b) *Señal escalda***

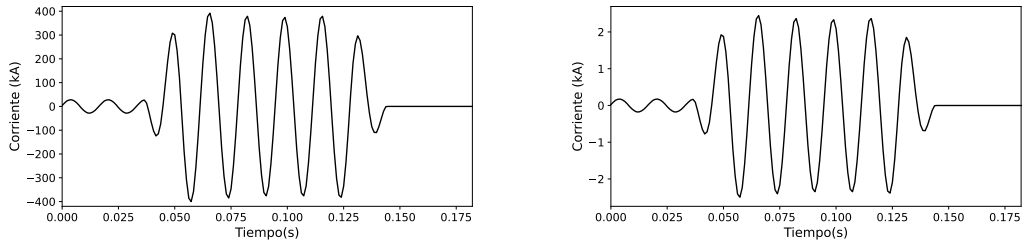
Figura 2.21: Comparativa entre una señal no escalada y una señal escalada.

2.5.2. Estandarización de los datos

La normalización de un conjunto de datos tiene como función centrar los datos con $\mu = 0$ y $\sigma = 1$. Los datos normalizados $\boldsymbol{z}$ suelen tener mejor resultado si son usados en modelos lineales o métodos de gradiente. Este calculo lo podemos realizar con la Ecuación (2.23).

$$\boldsymbol{z} = \frac{\mathbf{x} - \mu_x}{\sigma_x} \tag{2.23}$$

donde $\boldsymbol{x}$ es la señal de entrada , μ_x es la media de la señal de entrada y σ_x es la desviación estándar de la señal. El efecto de este escalado se muestra en la Figura 2.22.



(a) ***Señal original*** **(b)** ***Señal normalizada***

Figura 2.22: Estandarización de una señal

2.5.3. Normalización vectorial

La normalización es el proceso de escalar muestras individuales para tener una norma unitaria, el cual, suele ser útil cuando se trata de modelos de clasificación basados en distancia. Para normalizar de un conjunto de datos de dimensión j se hace uso de la Ecuación (2.24).

$$\mathbf{x_{norm}} = \frac{\mathbf{x}}{\left(\sum_{k=1}^{j} |\mathbf{x}_k|^p\right)^{\frac{1}{p}}} \tag{2.24}$$

para $p \in \mathbb{R}$, $p \geq 1$

Con p=2 se le conoce como la norma L^2 o como la norma euclidiana. Los valores mas usados de p para normalizar datos es 1 y 2.

2.6. Ingeniería de características.

Las aplicaciones basadas en ML están basadas en datos para producir modelos. Dichos datos poseen características implícitas que pueden ser derivadas de los mismos. Estos pueden ser, entre otros, de tipo estadístico (e.g., varianzas, medias) o transformaciones (e.g., Fourier, Wavelets). Estas características derivadas pueden utilizarse para incrementar el tamaño del espacio de características, lo cual proporcionará un mecanismo que aumente la

precisión del agrupamiento, y en consecuencia, la precisión de clasificación. En algunos otros casos tener muchas características puede derivar en clasificadores poco precisos, por lo que es necesario llevar a cabo una etapa de selección de características. La cual consiste en remover aquellas características que son poco útiles para la clasificación [Ikram Sumaiya Thaseen17]. En cuanto a los datos de clases múltiples, la prueba Chi-cuadrado ha demostrado ser efectiva en la selección de características [Yang97].

Adicionalmente, en el caso que se toca en este trabajo, nos permitirá hacer el agrupamiento con características simples aumentadas con algunas características derivadas de dominio de estudio, en este caso componentes simétricas aplicadas a fallas en lineas de distribución.

2.7. Resumen del cápitulo

En este capitulo se realizó una revisión de temas que serán necesarios para mejorar comprensión de la propuesta que sera desarrollada en los siguientes capítulos. Se ha expuesto la conformación básica de la un sistemas eléctrico de potencia, dándole mayor enfoque a las lineas de transición y el análisis de los 11 tipos de fallas que se pueden presentar en ellas. Además se detallo como se almacenan los eventos de falla en los archivos de formato COMTRADE. Después se realizó una revisión del algoritmo de k-Medias y las consideraciones a tomar para poderlo implementar. Y finalmente, se describe en que consiste la ingeniería de características.

Capítulo 3

Análisis y preprocesamiento de los datos

En este trabajo de tesis se ha implementado un sistema basado en k-Medias para la identificación y etiquetado de eventos de fallas en líneas de transmisión eléctrica dado su tipo de falla y que haga posible la clasificación de las mismas. Una etapa fundamental para la implementación de modelos de ML es el análisis de los datos que permitan convertirlos en datos útiles y de calidad. Lo anterior se lleva a cabo por medio de un preprocesamiento de los datos, el cual consiste en varios procesos. El primero en una fase de limpieza que nos den datos correctos, el segundo en la normalización, el tercero en la transformación y la integración de los datos, el cuarto en la imputación de los valores perdidos, el quinto en identificación de ruido y en algunos casos, un último que es la reducción de los datos. Por lo anterior, en este capítulo se analizan los registros de mediciones de voltaje y corriente almacenados en formato COMTRADE durante un evento de falla, que posteriormente serán convertidos en un conjunto de datos que induzca a un mejor agrupamiento al inicialmente encontrado haciendo uso de métodos de preprocesamiento de datos. Los conjuntos de datos obtenidos serán utilizados por modelos de k-Medias, expuestos en el capítulo siguiente, para agrupar los eventos. Posterior a ello se hace una evaluación del agrupamiento utilizando diversos clasificadores.

3.1. Descripción del conjunto de eventos

Se ha obtenido un conjunto de 107 eventos de falla almacenados en formato COMTRADE que proviene de un sistema de monitoreo de una red trifásica de potencia de generación centralizada. De antemano, se conoce la siguiente información:

1. Los eventos ocurrieron entre los años 2001 y 2005.
2. Provienen de 11 líneas de transmisión distintas.
3. Las líneas son de distintos voltajes de transmisión.
4. La frecuencia nominal de la red trifásica es de 60 Hz.
5. El tipo de falla de los eventos son por cortocircuito en las líneas de transmisión.
6. Algunos eventos activan el sistema de protección y ejecutan la desconexión de la misma, tal condición queda registrada como:
   (a) **TRIP**: son los eventos que causaron que los sistemas de protección desconectaran la línea de transmisión.
   (b) **NT**: son los eventos que no generaron una interrupción en la transmisión de energía eléctrica.
7. Las mediciones fueron registradas en 3 etapas que son: antes de la falla, durante de la falla y posterior a la falla.
8. Las mediciones correspondientes a las fallas tienen una duración de máximo 8 ciclos.

En la Tabla 3.1 se muestra una distribución del conjunto de los 107 eventos dividida por año de ocurrencia y su registro TRIP y NT. Podemos notar que se cuenta con 40 eventos TRIP y 67 eventos NT. También que los eventos son más recurrentes en algunas líneas que en otras, por ejemplo, de la línea L39U solo se cuenta con 1 evento, mientras que de la línea P16M se cuenta con 26 eventos registrados.

Tabla 3.1: Distribución del conjunto de eventos

LÍNEA	**2001**		**2002**		**2003**		**2004**		**2005**		***TOTAL***
	TRIP	**NT**	**TRIP**	**NT**	**TRIP**	**NT**	**TRIP**	**NT**	**TRIP**	**NT**	
L39U								1			1
L44M						2	2	5		1	10
P12C						3	1		11	6	21
P16M								1			1
P45S								3		2	5
P46L	1			5	1		6	6	4	3	26
S16M							2			7	9
S45P							3	6		3	12
Z24Z								6			6
Z65Z						3	2	2			7
Z000					1		4	1	2	1	9
TOTAL											**107**

Como se mencionó en la sección 2.3, el formato COMTRADE incluye un archivo .dat que contiene las mediciones de algunos parámetros de las líneas de transmisión. Para cada evento, el archivo .dat almacenado contiene las mediciones de voltaje y corriente de las 3 fases de la línea, así como también la estampa de tiempo del instante en que fueron tomadas las mediciones. Para leer cada uno de los archivos se ha usado la librería Pandas desarrollada para Python, la cual se especializa en el manejo y análisis de datos representados en tablas.

El resultado obtenido después de leer uno de los archivos .dat se puede observar en la Tabla 3.1, la cual cuenta con 7 columnas. La primera columna nombrada **t** contiene el tiempo en el que se tomaron las mediciones comenzando desde $\mathbf{t} = 0$, en las siguientes 3 columnas nombradas como **Ia**, **Ib** e **Ic** representan las mediciones del corriente, de la fase a, de la fase b y de la fase c respectivamente y las últimas 3 columnas nombradas como **Va**, **Vb** y **Vc** son las mediciones de voltaje en las fases a, b y c.

Tabla 3.2: Resultado de la lectura de un archivo .dat

t (s)	Ia (kA)	Ib (kA)	Ic (kA)	Va (kV)	Vb (kV)	Vc (kV)
0.000000	51.99620	-21.999900	-30.003600	-52.0001	63.2971	-11.3015
0.004166	-5.00211	47.000400	-43.002900	-43.2002	-23.4010	66.4966
0.008333	-52.00060	22.000300	29.997600	51.8993	-63.3000	11.3980
0.012500	4.99758	-48.000000	42.996900	43.1993	23.3980	-66.5001
...	...	...	...	...	...	...
0.166666	-0.00219	0.000192	-0.003882	-51.0001	64.2970	-13.2014
0.170833	-0.00219	0.000192	-0.003882	-44.8002	-21.7010	66.5965
0.175000	-0.00219	0.000192	-0.003882	51.0993	-64.4000	13.0979
0.179166	-0.00219	0.000192	-0.003882	44.7993	21.7980	-66.6000

Las mediciones contenidas en la Tabla 3.1 de voltaje y corriente se pueden observar en los gráficos de la Figura 3.1, donde se muestran las 6 gráficas relacionadas. Las 3 gráficas superiores corresponden a mediciones de corriente, mientras que las 3 gráficas inferiores corresponden a las mediciones de voltaje.

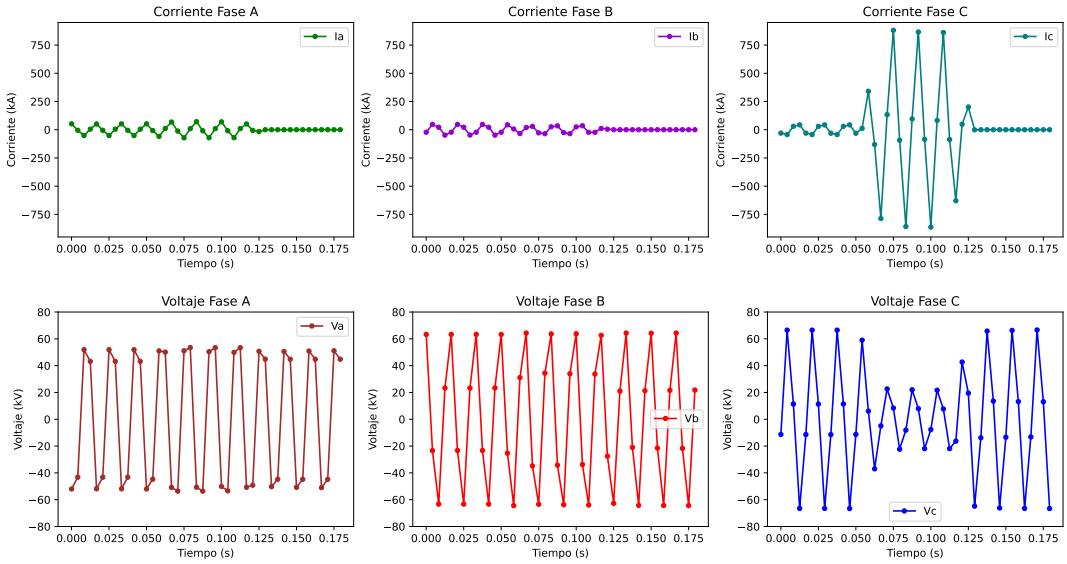


Figura 3.1: Gráficas de corrientes y voltajes

Como se puede notar, las mediciones de voltaje y corriente pueden ser tratadas como señales discretas, lo cual servirá para el análisis que se realizaran más adelante.

3.1.1. Análisis del conjunto de archivos

Para conocer más sobre los eventos se realizó el proceso de lectura para cada uno de los 107 archivos .dat, generando cada uno de ellos una tabla como la Tabla 3.1. Este grupo de tablas se han procesado para determinar el número de filas con las que cuentan, la tasas de muestreo y el número de ciclos muestreados. Un resumen de este procesamiento se puede observar en la Tabla 3.3, de la cual podemos determinar lo siguiente:

1. La tasa de muestreo es de 240 Hz y de 960 Hz.
2. Varía la duración del evento aún tomado de la misma línea.
3. Para la misma línea puede variar la tasa de muestreo y ciclos almacenados.
4. No todos los eventos cuentan con la misma cantidad de ciclos muestreados.

Tabla 3.3: Información del número de filas, la tasa de muestreo y la cantidad de ciclos muestreados para el conjunto de eventos

LÍNEA	Númer de filas								Tasa de muestreo		Ciclos muestreados						Total por línea
	44	60	160	176	192	208	240	256	240 Hz	960 Hz	10	11	12	13	15	16	
L39U					1					1			1				**1**
L44M					10					10			10				**10**
P12C		4	2			2	11	2	4	17	2			2	15	2	**21**
P16M	1								1			1					**1**
P45S	1			2	2				1	4		3	2				**5**
P46L	18			6	2				18	8		24	2				**26**
S16M	2			7					2	7		9					**9**
S45P	6			4	2				6	6		10	2				**12**
Z000		1					8		1	8					9		**9**
Z24Z	5			1					5	1		6					**6**
Z65Z	7								7			7					**7**
TOTAL	**107**								**107**		**107**						**107**

3.1.2. Análisis empírico de los eventos y etiquetado

El análisis empírico de los eventos consiste en examinar una serie de gráficas generadas a partir de los eventos almacenados por personal técnico especialista en el tema, quienes consideran los principios teóricos de análisis de fallas para hacer etiquetado de los eventos. Las variables eléctricas más observadas de cada evento son las mediciones de voltajes y corrientes de cada fase, así como también sus componentes simétricas. Adicionalmente,

también tienen a la mano información complementaria, en archivos adicionales, provista por el sistema de adquisición de información, como pueden ser contadores del tiempo de la falla, momento en el que se activa el sistema de protección y fase que inicia la falla.

Para explicar en que consiste el análisis empírico, usando solo la información con la que contamos, observemos la Figura 3.2, en la cual se puede observar en modo gráfico las mediciones de las corrientes y voltajes de las fases a, b y c respecto del tiempo. Se aprecia en el gráfico de la corriente de la fase b que los valores de corriente son muy altos en comparación con las corrientes de fase a y c. Adicional de ello, se observa un descenso del voltaje en la fase b. Hasta ahí se puede suponer que se cumple con las condiciones de falla del tipo monofásica BT, mencionadas en la sección 2.2.2.

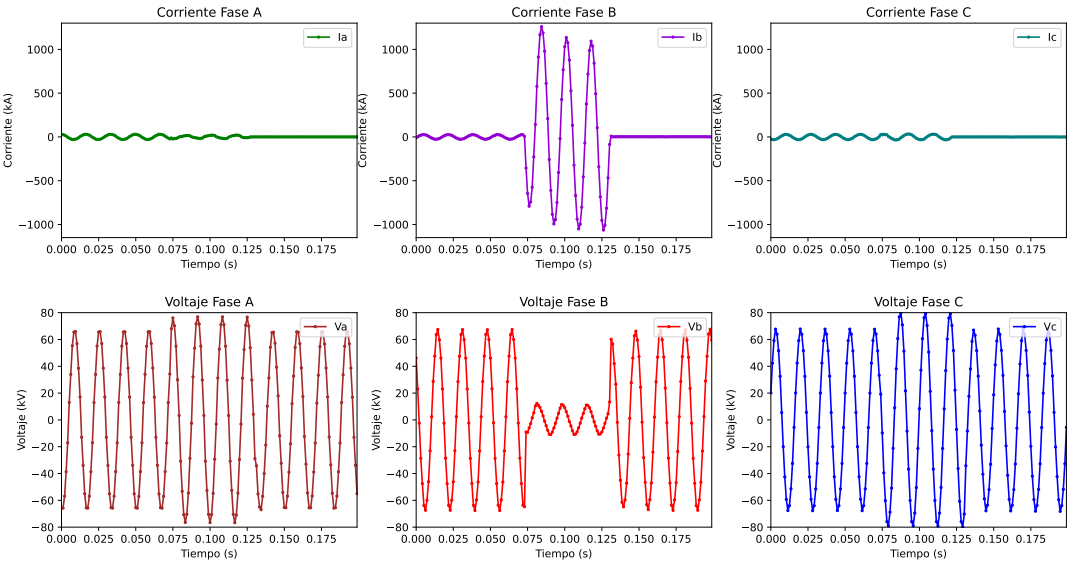


Figura 3.2: Mediciones de voltaje y corriente de una falla BT

Por otro lado, las magnitudes de las componentes simétricas se muestran en la Figura 3.3. Se aprecia en la imagen que las 3 componentes simétricas, durante tiempo de FALLA, son muy cercanas en magnitud. Si consideramos la evaluación de las componentes simétricas para fallas monofásicas, tenemos que $I_0 = I_1 = I_2$, se puede concluir que este evento es una falla monofásica del tipo BT. Cabe mencionar que para este ejemplo las componentes simétricas no son exactamente iguales, condición esperada para fallas monofásicas,

debido a que las corrientes de las fases a y c no son iguales a cero.

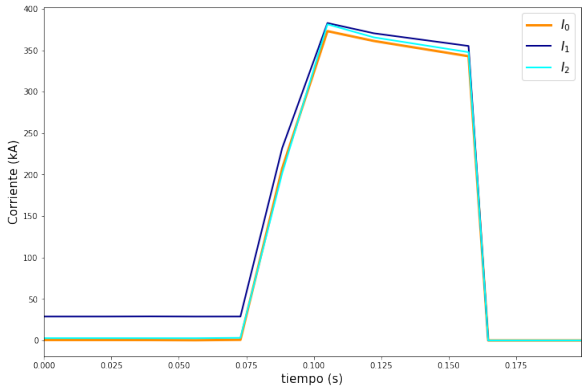


Figura 3.3: Componentes simétricas de una falla BT

Asesorado por un experto en líneas de transmisión y comparando con los principios para identificar los 11 tipos de fallas, que fueron presentados en el capítulo 2, se han etiquetado el conjunto de eventos usando el análisis empírico para etiquetar el conjunto de eventos en dos posibles conjuntos de etiquetas.

El primero es un etiquetado en 7 tipos de fallas que son AT, BT, CT, AB, AC, BC y ABC. Este conjunto de etiquetas serán conocidas en lo siguiente como $\mathcal{F}$. Se ha hecho este etiquetado inicial debido a la semejanza que guardan las fallas bifásicas y trifásicas con sus respectivas aterrizadas. En la Tabla 3.4 se muestra la cantidad de eventos etiquetados por tipo de falla. Mostrando mayor cantidad en las fallas monofásicas, seguidas por las bifásicas y las que menos hay son trifásicas.

Tabla 3.4: Etiquetado de eventos en 7 tipos de fallas

Tipo de falla	AT	BT	CT	AB	AC	BC	ABC
Cantidad	36	25	23	6	2	6	9

El segundo tipo de etiquetado, se realiza considerando los 11 tipos de fallas. El resultado obtenido se muestra en la Tabla 3.5, en la que destaca que del tipo AC no hay ningún evento.

Tabla 3.5: Etiquetado de eventos en 11 tipos de fallas

Tipo de falla	AT	BT	CT	AB	AC	BC	ABT	ACT	BCT	ABC	$ABCT$
Cantidad	36	25	23	3	0	1	3	2	5	1	8

Entonces, las etiquetas resultantes por el análisis empírico para 11 tipos de falla son: AT, BT, CT, AB, AC, BC, ABT, ACT, BCT, ABC, $ABCT$, a este conjunto de etiquetas se les conocerá como $\mathcal{G}$.

El análisis empírico suele ser tardado y el resultado del análisis puede variar dependiendo del personal técnico que lo haya realizado, de la calidad de las mediciones y de la información con la que se cuente.

Como se observa en la Tabla 3.5 para el tipo de falla AC no hay eventos y para la ACT solo se cuenta con dos, por debajo de las demás fallas bifásicas. Este desbalance de cantidad de eventos afectaría significativamente a los modelos propuestos para clasificación, es por ello que se han generado dos eventos sintéticos basados en fallas del tipo AB y ABT, tales que no afecten el funcionamiento del sistema propuesto. Estos dos eventos serán agregados al conjunto de eventos que será procesado.

3.2. Transformación del conjunto de eventos en conjuntos de datos

La información almacenada en los archivos .dat de cada uno de los eventos ha permitido etiquetar los eventos por tipo falla. Esto se ha logrado mediante un análisis de las gráficas generadas a partir de las mediciones almacenadas y de su interpretación por un experto en el tema, tal como se describió en la sección anterior.

La tarea de etiquetar eventos se realizará de manera automática haciendo uso de herramientas de ML. Por lo tanto, se puede decir que la problemática que se desea resolver es dividir el conjunto de eventos en 11 posibles grupos que representan a cada una de las 11 posibles tipos de fallas. Entonces esta tarea se convierte en un problema de agrupamiento, ya que es intentar agrupar eventos que comparten características semejantes. Por lo cual, se propone usar el método de agrupamiento k-Medias como herramienta para

evaluar el agrupamiento obtenido atreves del procesamiento de datos en diferentes etapas de procesamiento y proponer, con base en el resultado, un etiquetado posible por tipo de falla.

Para poder implementar el algoritmo k-Medias es necesario definir que información de la que se cuenta es de utilidad y que otra podemos agregar para lograr un buen agrupamiento. Para ello se ha determinado inicialmente evaluar la calidad de la información, lo que consiste en revisar si la información está completa, es consistente, es precisa y/o es coherente. Si la calidad de la información no es buena, se puede mejorar aplicando algunas técnicas de pro-procesamiento de datos para lograrlo.

A continuación se muestran las técnicas que se han implementado para pasar de un conjunto de eventos a un conjunto de datos que serán sometidos aún agrupamiento mediante un algoritmo de k-Medias, dando mayor importancia a la mejora de la calidad de los datos.

3.2.1. Conversión a una tasa única de muestreo

Como se observó en la Tabla 3.3, la tasa se muestreó no fue la misma que se utilizó para tomar las mediciones almacenadas en los archivos .dat de cada uno de los eventos y puede ser de 240 Hz o 960 Hz, lo que representa las armónicas 4 y 16 respectivamente. Se pretende hacer nuevos archivos con una única tasa de muestreo basados en los archivos .dat de los eventos. Por lo anterior, se ha considerado que las mediciones no contienen ruido, están filtradas, no tienen datos faltantes y solo se desea trabajar con la componente fundamental de la frecuencia de la señal muestreada. Asumiendo lo anterior, se hace el cálculo mínimo de la frecuencia de muestreo necesaria aplicando la Ecuación (2.15), donde f_m será igual a la frecuencia nominal de la línea(60 Hz) más un 10 %, considerando de una variación máxima permisible. Por lo tanto, f_m=66 Hz y la tasa de muestreo es:

$$\begin{aligned} f_s &\geq 2f_m \\ &\geq 2(66Hz) \\ &\geq 132Hz \end{aligned}$$

Por lo anterior, la tasa de muestreo mínima deberá ser de al menos de 132 Hz. Debido a que tanto la frecuencia de 240 Hz y la de 960 Hz son mayor a f_s, se ha seleccionado como la tasa de muestreo única a 240 Hz, que para efectos de este trabajo es de utilidad, ya que es la tasa más cercana al valor f_s que cuenta el conjunto de eventos y no será necesario hacer un sobremuestreo a diferencia de haber sido seleccionada la frecuencia de 960 Hz.

Para hacer el proceso de transformar a una tasa única de muestreo, los archivos que estaban generados a una tasa de 960 Hz se aplicó una técnica de sub-muestreo que deja 1 de cada 4 mediciones. Mientras que los archivos generados a una tasa de muestreo de 240 Hz no se les realizó ningún cambio.

Cabe aclarar que durante el evento de falla puede haber una modificación en la frecuencia, sin embargo, para sistemas de potencia en donde los generadores proveen una inercia muy grande las variaciones son muy pequeñas. Si nuestra falla ocurriera en un sistema con un generador aislado, entonces, en este caso, la frecuencia si variaría significativamente por no haber la suficiente inercia de los generadores. Por otro lado, también se entiende que existen componentes de alta frecuencia en las redes de transmisión, sin embargo, no afectan de manera significativa el modelo propuesto en él presenten trabajo y esto quedara claro en capítulos posteriores.

3.2.2. Segmentación del evento

Se conoce que las mediciones almacenadas durante el evento fueron tomadas antes que comenzara la falla(PRE-FALLA), durante la falla(FALLA) y posterior a la falla(POS-FALLA). Se conoce de antemano que la mayoría de las fallas cuentan con 8 ciclos almacenados de la etapa de FALLA, mientras que de las otras etapas se desconoce cuántos ciclos están almacenados de cada uno.

Se propone solo trabajar con las mediciones corresponden a la etapa de la FALLA, considerando una duración de 8 ciclos de falla. Dado que del proceso anterior, los archivos .dat tienen una tasa de muestreo única, la cual es de 240 Hz, lo que es igual a 4 muestras por ciclo, se concluye que al separar las mediciones correspondientes de la FALLA resultarán con 32 mediciones por parámetro almacenado.

A continuación se describen las acciones que fueron tomadas para llevar a cabo

la tarea de aislar las mediciones de FALLA, teniendo en cuenta que se obtendrán nuevos archivos .dat que contengan cada uno 32 mediciones de corriente por cada fase.

Detección del inicio de la FALLA

Considerando solo las mediciones de corriente de cada una de las fases, se sabe que la FALLA comenzará con un cambio abrupto de la magnitud de la corriente. Para detectar inicio de la FALLA, primeramente se han tomado en cuenta los primeros 3 ciclos de mediciones de PRE-FALLA para determinar la máxima medición absoluta de la corriente por cada una de las fases. Posterior a ello, se ha buscado una medición tal, que su valor absoluto sea mayor al 20 % de la máxima medición absoluta, en secuencia del tiempo. Donde se cumple dicha condición se considera que inicia la falla.

Selección de mediciones posteriores al inicio de la falla

Posterior a la detección del inicio de la falla, se seleccionaron los siguientes 8 ciclos, donde se considera que se encuentra la FALLA. Sin embargo, en algunos casos la zona de FALLA detectada dura menos de 8 ciclos. Por lo que, para completarlos, se repitió el último ciclo tantas veces fuera necesario. Esta decisión es tomada en el entendido que las mediciones correspondientes a la FALLA únicamente tendrán en cuenta la información de la falla, y en el caso de datos faltantes se asume con esta repetición de mediciones que la falla solamente continuó sin finalizar.

Este proceso se ha realizado a todos los archivos .dat y se ha hecho un nuevo conjunto de archivos, solo con las mediciones de corriente resultantes, cuya principal característica es tener 32 mediciones de corriente de cada fase. A este conjunto de archivos les llamaremos AF_1.

Se ha sometido el conjunto AF_1 a una evaluación al agrupamiento usando iVAT siguiendo el siguiente proceso:

1. Se realizó la lectura de cada uno de los archivos en AF_1.
2. Cada archivo leído se almacenó en una matriz $\mathbf{F}$ de 32 filas por 3 columnas, donde las columnas son las mediciones de corrientes de las 3 fases.

3. Cada matriz **F** fue trasladada a una matriz **X** de una fila por 96 columnas. Las primeras 32 columnas son las mediciones de corriente de la fase a, las siguientes 32 son las mediciones de corriente de la fase b y las últimas 32 columnas son las mediciones de corriente de la fase c

4. Las 109 matrices resultantes se concatenaron para formar una matriz **D** de dimensión 109×96.

5. Se aplicó el Algoritmo 3 a la matriz $\mathbf{D}_1$

El resultado obtenido de realizar el proceso anterior está representado en la Figura 3.4. Se puede notar que no existe una tendencia clara al agrupamiento. A partir de ahora, llamaremos, a la matriz $\mathbf{D}_1$ conjunto de datos X_1

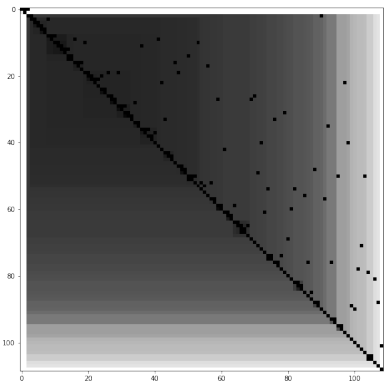


Figura 3.4: iVAT aplicado a X_1

Eliminación de mediciones tomadas durante la activación del sistema de protección

Para intentar mejorar la calidad de los datos e inducir a mejor agrupamiento, se realizó una inspección de las mediciones contenidas en los archivos en conjunto AF_1 y se determinó que también incluyen mediciones de POS-FALLA, que corresponden a las mediciones tomadas mientras el sistema de protección se encuentra activo.

Por lo que se propuso eliminar esta zona de POS-FALLA, caracterizada por tener mediciones de valor constante cercanas a cero. Cada uno de los archivos en AF_1 fueron procesados para detectar de manera automática la presencia de esta POS-FALLA y fue reemplazada por el último ciclo de la zona de FALLA, tantas veces fuera necesario para completar los 8 ciclos, generando un nuevo conjunto de archivos a los que llamaremos AF_2.

El conjunto AF_2 fue sometido a un proceso semejante al de la sección para obtener una nueva matriz $\mathbf{D}$ a la cual nombraremos ahora $\mathbf{D}_2$, que al aplicarle el Algoritmo 3 se obtiene la Figura 3.5. Se puede notar que existen zonas más oscuras que antes no se tenían, lo que se interpreta como un mejor agrupamiento, sin embargo, siguen sin observarse un agrupamiento claro. La matriz $\mathbf{D}_2$ será conocida como conjunto de datos X_2.

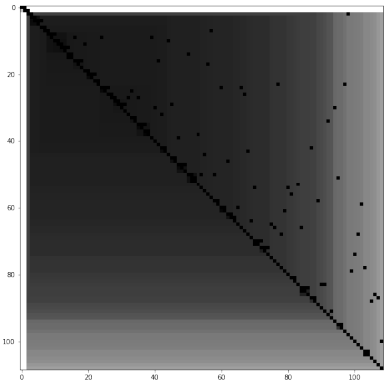
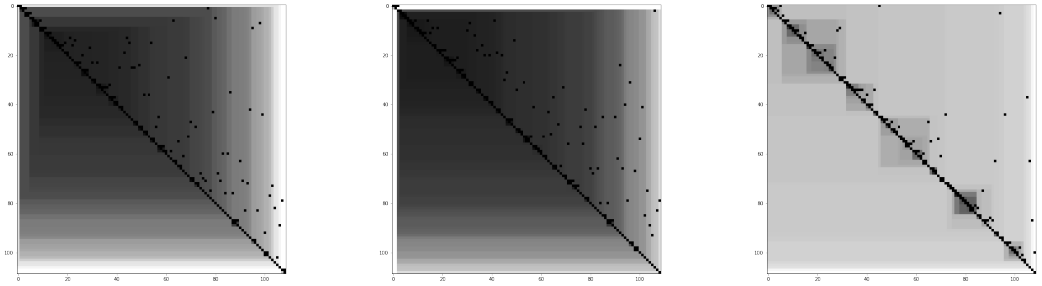


Figura 3.5: iVAT aplicado a X_2

3.3. Transformación del conjunto datos

Se han aplicado las técnicas de preprocesamiento de datos como son el escalado entre 0 y 1, la estandarización y la normalización vectorial a los conjuntos de datos X_2, haciendo 3 nuevos conjuntos datos que pueden ser evaluados usando iVAT. En la Figura 3.6 se muestra el resultado de la transformación y el escalado a los conjuntos de datos X_2. Se puede notar que normalización vectorial muestra grupos más claros sobre la diagonal, lo cual se interpreta que el conjunto de datos normalizado mejoró la calidad del agrupamiento,

sin embargo, hay destacar que los grupos no se observan muy oscuros para considerar una buena calidad del agrupamiento.



(a) ***Escalado entre 0 y 1*** **(b)** ***Estandarización*** **(c)** ***Normalización vectorial***

Figura 3.6: iVAT al aplicar transformación y escalado del conjunto X_2

Al conjunto de datos X_2 normalizado se le conocerá como X_3.

3.4. Aplicación de ingeniería de características

Se ha aplicado ingeniería de características al conjunto de archivos FA_2, procesando cada uno de ellos para extraer ciertas características de las señales de corriente. Las características extraídas están relacionadas con medidas estadísticas y medidas utilizadas en el estudio de señales eléctricas. A continuación, se detallan las siguientes características extraídas:

1. La media ($\overline{\boldsymbol{I_a}}$,$\overline{\boldsymbol{I_b}}$,$\overline{\boldsymbol{I_c}}$).
2. El valor mínimo($min(I_a)$,$min(I_b)$, $min(I_c)$).
3. El valor máximo($max(I_a)$,$max(I_b)$, $max(I_c)$).
4. Los cuartiles 1,2 y 3 ($Q_1(I_a)$,$Q_1(I_b)$,$Q_1(I_c)$,$Q_2(I_a)$,$Q_2(I_b)$,$Q_2(I_c)$,$Q_3(I_a)$,$Q_3(I_b)$,$Q_3(I_c)$).
5. La varianza($\sigma^2(I_a)$,$\sigma^2(I_b)$,$\sigma^2(I_c)$).

6. La desviación estándar ($\sigma(I_a)$,$\sigma(I_b)$,$\sigma(I_c)$).

7. El valor RMS($RMS(I_a)$,$RMS(I_b)$,$RMS(I_c)$).

8. La varianza de la suma de corrientes.

Por lo que cada uno de los archivos quedo representado por un vector de características de 28 dimensiones, que también se puede representar por una matriz dé 1×28. Al igual que en las secciones pasadas, las 109 matrices se pueden integrar en una sola matriz de 109×28. Posteriormente a esta matriz se le ha realizado la prueba Chi-cuadrado para hacer una selección de características, de la cual resultó que las características más útiles son las varianzas de las corrientes. Considerando este resultado se ha formado un nuevo conjunto de datos con las varianzas de las mediciones de corriente de cada fase que después se ha normalizado, dada la mejora que ha traído al agrupamiento la normalización en la sección anterior.

Por lo tanto, el conjunto de datos que se ha generado de a partir de las varianzas de X_1 y ha sido normalizado se le conocerá como el conjunto de datos X_4.

En la Figura 3.7 se muestran el mapa de color de los conjuntos de datos X_4 al aplicar iVAT. Se observa que la calidad en el agrupamiento ha mejorado muy significativamente respecto al conjunto de datos inicial X_1.

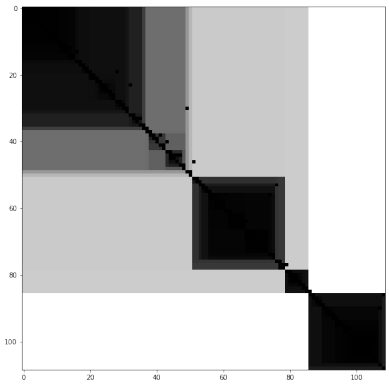


Figura 3.7: iVAT aplicado a X_4

3.5. Resumen del cápitulo

Se ha comenzado el capítulo describiendo la información de 107 eventos almacenados en formato COMTRADE. Del análisis se ha determinado que los archivo .dat contiene el registro de mediciones de voltaje y corriente, así como también, una estampa de tiempo del evento. Después, los archivos se procesaron para realizar un análisis empírico de la información, de la cual, han resultado dos conjuntos de etiquetados por tipo de falla. El primer etiquetado corresponde al análisis para 7 tipos de fallas nombradas $\mathcal{F}$ y el segundo etiquetado corresponde al análisis para 11 tipos de fallas nombradas $\mathcal{G}$.

Posteriormente, todos los archivo .dat han sido nuevamente procesados y analizados para obtener la información contenida en ellos y convertirla en datos que forman un primer conjunto de datos de 96 dimensiones llamado X_1. Después X_1 ha sido procesado, usando una serie de técnicas de preprocesamiento de datos, para mejorar su tendencia al agrupamiento, resultando de cada técnica utilizada un nuevo conjunto de datos. El total de conjuntos resultantes son: X_1, X_2, X_3, X_4. A los 4 conjuntos de datos se le conocerá como el conjunto $\mathcal{X}$. Cabe mencionar que el grupo mejor evaluado ha sido X_4 contando con tan solo 3 dimensiones.

Capítulo 4

Implementación y evaluación de los modelos

En este capítulo, se describe la implementación del modelo propuesto para agrupar los conjuntos de datos en $\mathcal{X}$, en 11 grupos que representan los 11 tipos de fallas. Dicho proceso se realizó en dos etapas. La primera etapa consiste en agrupar cada conjunto de $\mathcal{X}$ en 7 grupos, usando el algoritmo de agrupamiento k-Medias. Cada agrupamiento resultante será comparado con el etiquetado $\mathcal{F}$, descrito en la sección 3.1.2, con la finalidad definir el mejor conjunto de datos en $\mathcal{X}$. En la segunda etapa, se agrupa el conjunto de datos X_4 en 11 grupos, agregando una característica derivada de componentes simétricas y usando el algoritmo k-Medias. Posteriormente, se realiza una comparación de los agrupamientos obtenidos con las etiquetas $\mathcal{G}$ y se discuten los resultados obtenidos.

4.1. Definición del conjunto de entrenamiento y de prueba

Cada conjunto en $\mathcal{X}$ ha sido separado en un conjunto entrenamiento y un conjunto de pruebas, con el propósito de que el modelo generado con el conjunto de entrenamiento sea evaluado con un conjunto de datos desconocido para el mismo. El conjunto de entrenamiento será utilizado para realizar el agrupamiento, mientras que el conjunto de prueba será utilizado para evaluar a los clasificadores. Por lo tanto, el conjunto de entrenamien-

to de $\mathcal{X}$ quedara definido como $\mathcal{X}^E$ =$\{X_1^E,\ X_2^E, X_3^E, X_4^E\}$, y conjunto de pruebas como $\mathcal{X}^P = \{X_1^E, X_2^E, X_3^E, X_4^E\}$.

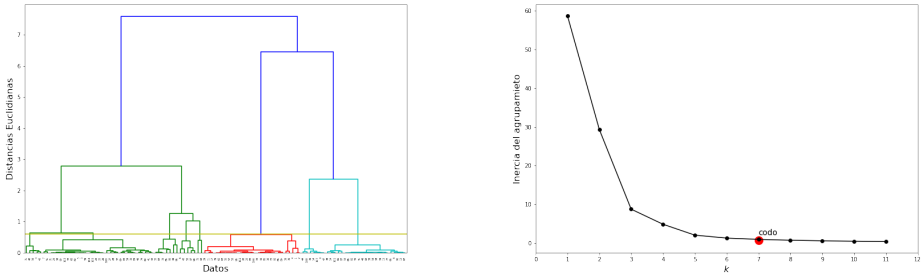
Se han seleccionado el 80 % de los datos como conjunto de entrenamiento y el 20 % como conjunto de prueba para cada conjunto en $\mathcal{X}$. La Tabla 4.1 muestra la cantidad de datos tomados para el conjunto de prueba y para el de entrenamiento en función de $\mathcal{G}$, esto para cada conjunto en $\mathcal{X}$. De esta tabla se puede notar que de las etiquetas AC, BC y ABC no se tomó ningún dato, debido a que solo se cuenta con uno para estos tipos de fallas.

Tabla 4.1: Cantidad seleccionada de datos para entrenamiento y prueba

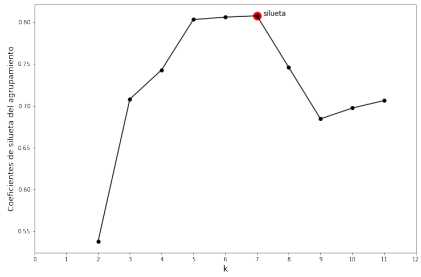
Etiquetas $\mathcal{G}$	AT	BT	CT	AB	AC	BC	ABT	ACT	BCT	ABC	$ABCT$
Conjunto de entrenamiento	30	21	19	2	1	1	2	2	4	1	6
Conjunto de pruebas	6	4	4	1	0	0	1	1	1	0	2

4.2. Evaluando el conjunto de entrenamiento para determinar el valor de k

Aunque se conoce de manera empírica la cantidad de grupos posibles del conjunto de eventos, no necesariamente el conjunto de datos generados del procesamiento que se le han dado a las mediciones derive en la misma cantidad de grupos, aunque si sea lo esperado. Por lo tanto, se han implementado el método del codo, el método de la silueta y dendrograma para tener una idea de la cantidad de grupos que pueden formar los conjuntos de datos. El conjunto de datos seleccionado es X_4, debido a que mostró una buena evaluación de la tendencia al agrupamiento.



(a) ***Dendrograma*** **(b)** ***Método del codo***



(c) ***Método de la silueta***

Figura 4.1: Métodos para determinar el valor de k

La aplicación de los métodos para determinar el valor de k son mostrados en la Figura 4.1 para X_4. De la cual, en la Figura 4.1(a) se observa un dendrograma de los datos usando la distancia euclidiana como métodos de medición. Se puede observar que el valor grupos k sugeridos por este método es 8. Mientras que en la Figura 4.1(b), se muestra el método del codo, en la cual se señala con un punto en color rojo el valor correspondiente a 7 grupos. En este punto es donde se observa el comienzo de una menor variación de cambio respecto al siguiente valor de inercia. Y en la Figura 4.1(c) se muestra el resultado del método de la silueta mostrando el máximo valor en $k = 7$. De acuerdo al análisis anterior, se decide comenzar con un valor dé $k = 7$.

4.3. Modelos de agrupamiento de 7 grupos

Se ha implementado el algoritmo de agrupamiento k-Medias con k=7 y con método de inicialización de centroides utilizado en k-Medias++. El resultado obtenido del agrupamiento ha sido comparado con las etiquetas $\mathcal{F}$ haciendo uso de la matriz de confusión. En las siguientes subsecciones se exponen los resultados para cada uno de los conjuntos en $\mathcal{X}$.

4.3.1. Modelado con los datos en X_1^E

Los datos en X_1^E fueron formados a partir de las mediciones de corriente de los archivos .dat del conjunto de eventos inicial, unificadas a una sola tasa de muestreo. Cada uno de los datos es de 96 dimensiones. El proceso para la obtención de este conjunto de datos se describió en la subsección 3.2.2.

Se ha ejecutado proceso de agrupamiento usando el conjunto X_1^E mediante la implementación del algoritmo de k-Medias con $k = 7$. El agrupamiento obtenido ha sido comparado con $\mathcal{F}$ generando una matriz de confusión, la cual se muestra en un mapa de color en la Figura 4.2. Las etiquetas $\mathcal{F}$ representan las columnas, mientras que el conjunto de etiquetas $\widehat{\mathcal{F}}$= {$\widehat{AT}$,$\widehat{BT}$, $\widehat{CT}$,$\widehat{AB}$,$\widehat{AC}$, $\widehat{BC}$,$\widehat{ABT}$,$\widehat{ACT}$, $\widehat{BCT}$,$\widehat{ABC}$, $\widehat{ABCT}$} son las filas, las cuales son el etiquetado que se le ha dado a cada grupo resultante del agrupamiento.

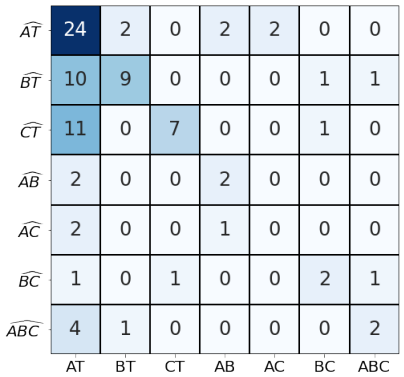


Figura 4.2: Matriz de confusión de $\widehat{\mathcal{F}}$ y $\mathcal{F}$ con X_1^E

Del análisis de la Figura 4.2 se concluye que los resultados verdaderos positivos

son 46 datos de 89 posibles.

4.3.2. Modelado con los datos en X_2^E

El conjunto de datos X_2^E han sido formados a partir de X_1^E, con la variación de que las mediciones tomadas después del disparo del los sistemas de protección se reconstruyeron con los mismos datos de la falla. El proceso para la obtención de este conjunto de datos se describió en la subsección 3.2.2.

Se ha ejecutado proceso de agrupamiento usando X_2^E mediante la implementación del algoritmo de k-Medias con $k = 7$. El agrupamiento obtenido ha sido comparado con $\mathcal{F}$ generando una matriz de confusión, la cual se muestra en un mapa de color en la Figura 4.3.

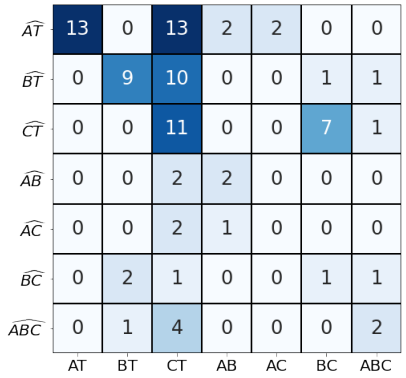


Figura 4.3: Matriz de confusión de $\widehat{\mathcal{F}}$ y $\mathcal{F}$ con X_2^E

Del análisis de la Figura 4.3 se concluye que los resultados verdaderos positivos son 38 datos de 89 datos.

4.3.3. Modelado con los datos en X_3^E

Los datos X_3^E es el resultado de la normalización de los datos X_2^E. Se ha implementado el proceso de agrupamiento usando el conjunto X_3^E mediante la implementación del algoritmo de k-Medias con $k = 7$. El proceso para la obtención de este conjunto de datos se describió en la sección 3.3. El proceso para la obtención de este conjunto de datos

se describió en la subsección 3.4.

El agrupamiento obtenido ha sido comparado con $\mathcal{F}$ generando una matriz de confusión, la cual se muestra en un mapa de color en la Figura 4.4.

Del análisis de la Figura 4.4 se concluye que los resultados verdaderos positivos son 38 datos de 89 datos.

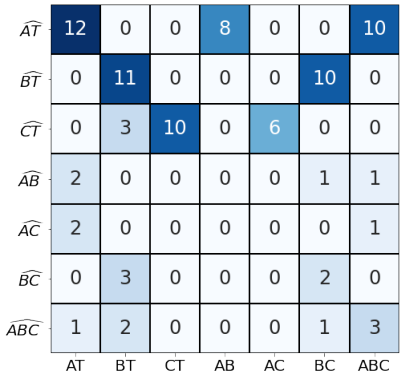


Figura 4.4: Matriz de confusión de $\widehat{\mathcal{F}}$ y $\mathcal{F}$ con X_3^E

4.3.4. Modelado con los datos en X_4^E

Los datos X_4^E son las varianzas de los datos en X_2^E normalizadas. Se ha ejecutado proceso de agrupamiento usando el conjunto X_4^E mediante la implementación del algoritmo de k-Medias con $k = 7$. El agrupamiento obtenido ha sido comparado con $\mathcal{F}$ generando una matriz de confusión, la cual se muestra en un mapa de color en la Figura 4.5.

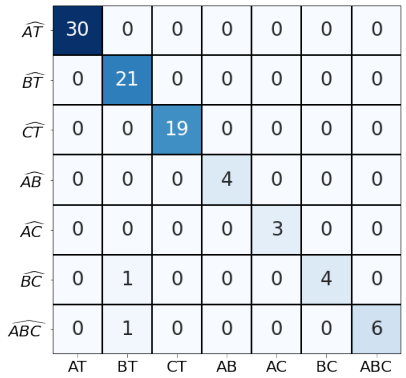


Figura 4.5: Matriz de confusión de $\widehat{\mathcal{F}}$ y $\mathcal{F}$ con X_4^E

4.3.5. Comparativa de modelos con 7 grupos

En Tabla 4.2 se reporta una evaluación adicional con las métricas de precisión, Exhaustividad y exactitud para cada uno de los agrupamientos de los conjuntos de datos en $\mathcal{X}$. De la cual destaca que el conjunto de datos X_4 muestra el mayor valor de verdaderos positivos con un valor de 87, teniendo un 0.99 de precisión de un máximo de 1, mientras que el menor es obtenido usando el X_2^E con una precisión de 0.42.

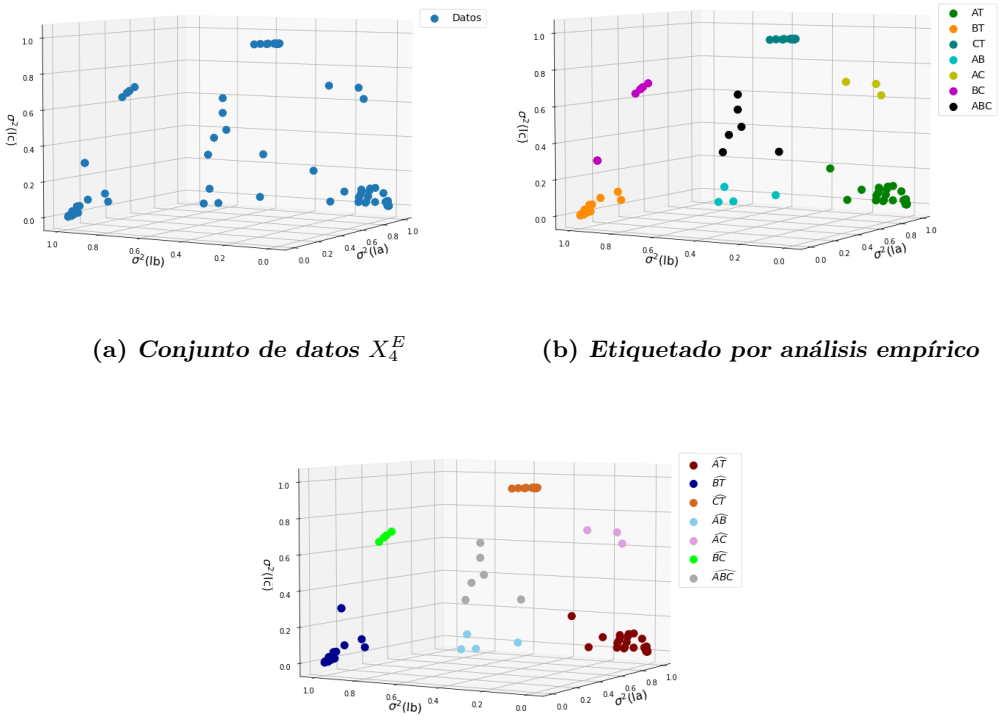
Tabla 4.2: Comparativa del agrupamiento con k-Medias y el etiquetado empírico

Conjunto datos	Verdadero/Positivo	Precisión	Exhaustividad	Exactitud	F1
X_1^E	46	0.5	0.4	0.52	0.41
X_2^E	38	0.42	0.35	0.43	0.35
X_3^E	38	0.38	0.33	0.43	0.32
X_4^E	87	0.99	0.95	0.89	0.97

Por lo tanto, se ha considerado que la forma de pre-procesar los datos, es usando las técnicas que han llevado a formar los conjuntos de datos X_4.

4.3.6. Representación gráfica del modelo seleccionado

El conjunto X_4^E se puede representar en un gráfico de 3 dimensiones donde cada eje representara a cada una de las varianzas de la corriente, tal como se muestra en la Figura 4.6(a). Usando estos datos también se puede hacer la representación de los conjuntos etiquetados por $\mathcal{F}$, como se observa en la Figura 4.6(b). Al realizar el agrupamiento, el resultado de igual manera puede ser representado en 3 dimensiones como se muestra en la Figura 4.6(c). Si realizamos un análisis de los grupos entre las Figuras 4.6(b) y 4.6(c), podemos notar la similitud que existe entre ambos agrupamientos.



(a) ***Conjunto de datos*** X_4^E **(b)** ***Etiquetado por análisis empírico***

(c) ***Agrupamiento***

Figura 4.6: Datos representados en 3 dimensiones

4.4. Modelos de agrupamiento de 11 grupos

4.4.1. Modelado con los datos en X_4^E

Se ha agrupado el conjunto X_4^E en 11 grupos usando k-Medias. Ahora se ha comparado el resultado con el agrupamiento formado por el etiquetado $\mathcal{G}$, formado la matriz de confusión que se muestra en el mapa de color en la Figura 4.7. En dicha imagen se puede observar que la cantidad de valores verdaderos positivos es de 64 de 89.

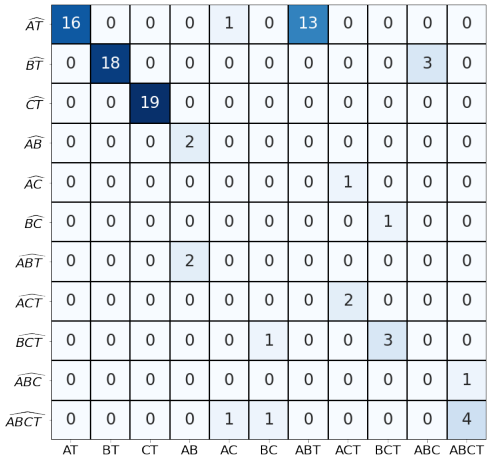
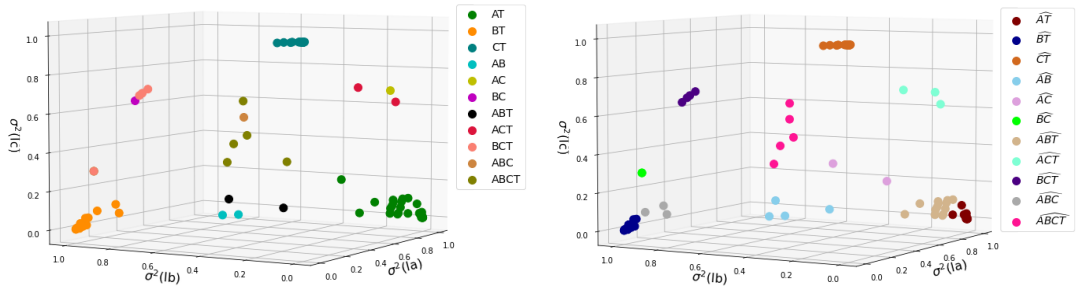


Figura 4.7: Matriz de confusión del agrupamiento obtenido X_4^E y los grupos etiquetados con $\mathcal{G}$

Otra manera de observarse el resultado del agrupamiento es de manera gráfica, tal como se muestra en la Figura 4.8(b), donde se puede ver que hay 11 grupos, sin embargo, si se compara con la Figura 4.8(a), la cual muestra una representación, el agrupamiento generado por las etiquetas $\mathcal{G}$, se nota que son pocas las similitudes entre los agrupamientos.



(a) ***Etiquetado por análisis empírico*** **(b)** ***Agrupamiento***

Figura 4.8: Datos representados en 3 dimensiones

Al observar la Figura 4.8(a) se puede notar que las etiquetas correspondientes a los tipos de fallas AB, AC, BC y ABC se encuentran muy cercanas a las etiquetas con tipos de fallas ABT, ACT, BCT y $ABCT$ respectivamente, lo cual no permite que se lleve a cabo el agrupamiento esperado. Entre algunas técnicas que podemos implementar para mejorar el resultado del agrupamiento, podemos mencionar agregar más característica que propicien un mejor agrupamiento en un espacio de mayor dimensionalidad o haciendo algún preprocesamiento o pos-procesamiento adicional los datos. Tal como se mostrara en las siguientes secciones.

4.4.2. Modelo usando componente simétrica de secuencia cero como característica

Para separar los datos que representan a las fallas bifásicas y trifásicas en aterrizadas y no aterrizadas, se ha propuesto agregar una cuarta característica que está en función del valor de la componente cero. Para determinar el valor de la cuarta característica para cada dato se calcula la componente de secuencia cero para las mediciones tomadas en PRE-FALLA y en FALLA. Si la magnitud de I_{a_0} en FALLA es mucho mayor que la que tenía en PRE-FALLA, se le asignará un 1 como cuarta característica y si no se asignará un 0. A este nuevo conjunto se le llamará X_5^E.

Se ha agrupado el conjunto X_5^E en 11 grupos usando k-Medias. Ahora se ha comparado el resultado con el agrupamiento formado por el etiquetado $\mathcal{G}$, formado la matriz de confusión que se muestra en el mapa de color en la Figura 4.9. En dicha imagen se puede observar que la cantidad de valores verdaderos positivos es de 83 de 89. La mejora respecto al resultado mostrado en la sección anterior es significativo, aunque no se iguala con el resultado obtenido en 4.3.4 que es lo esperado.

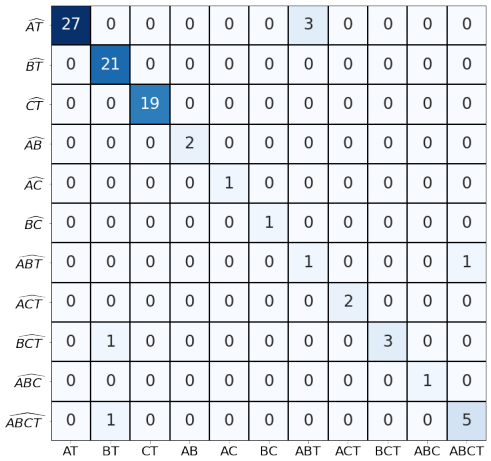


Figura 4.9: Matriz de confusión del agrupamiento obtenido X_5^E con componente simétrica y los grupos etiquetados con $\mathcal{G}$

4.4.3. Modelo usando componente simétrica de secuencia cero como pos-procesamiento

Para separar los datos que representan a las fallas bifásicas y trifásicas en aterrizadas y no aterrizadas, se usa la componente simétrica de secuencia cero como criterio de selección. La implementación se ha realizado de la siguiente manera:

1. Para los datos del conjunto X_4^E, se calcula el valor de su componente simétrica de secuencia cero.

2. Aquellos datos cuyo valor de la componente de secuencia cero es mucho mayor a la que contaba en la zona PRE-PREFALLA formaran parte de un nuevo grupo llamado

X_{F0}, y si no lo es, formará parte de X_{T0}.

3. Los datos en X_{F0} conservaran su etiqueta, mientras que los datos en X_{T0} que están etiquetados como $\widehat{AB}$,$\widehat{AC}$,$\widehat{BC}$,$\widehat{ABC}$ y $\widehat{AB}$ serán reemplazadas por $\widehat{ABT}$,$\widehat{ACT}$,$\widehat{BCT}$ y $\widehat{ABCT}$, respectivamente.

De lo anterior, se han obtenido 11 grupos. El agrupamiento dado de X_{T0} se considera que son fallas aterrizadas y el agrupamiento de X_{F0} se considera que son fallas no aterrizadas. Teniendo en cuenta esto, se realiza una comparación de estos 11 grupos con los grupos generados por las etiquetas $\mathcal{G}$. El resultado obtenido se muestra el mapa de color de la Figura 4.10. El valor verdadero positivo para esta comparación es de 87. Con este resultado se ha igualado el resultado obtenido en 4.3.4 que era lo esperado.

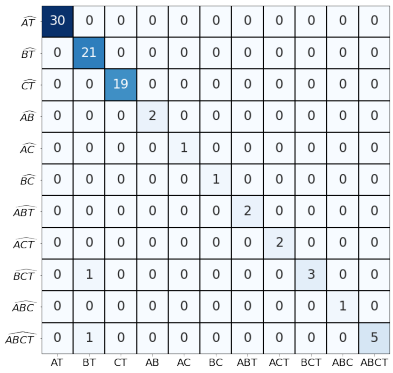


Figura 4.10: Matriz de confusión del agrupamiento obtenido de X_{T0} y X_{F0} con los grupos etiquetados con $\mathcal{F}_{11}$

Este agrupamiento se representa en la Figura 4.11, en un gráfico de 3 dimensiones. Como se puede notar, es muy parecido el agrupamiento con la Figura 4.8(a).

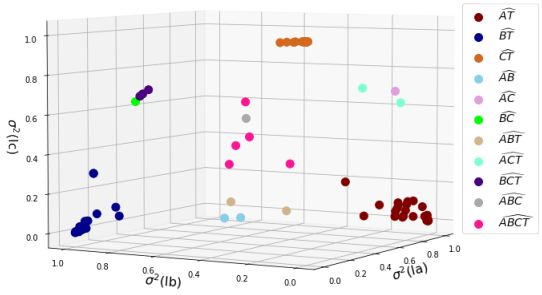


Figura 4.11: Agrupamiento en 11 grupos usando componentes simétricas de secuencia cero

4.5. Resumen del cápitulo

En este capítulo, se describe la proporción en la que se han separado cada conjunto de datos en $\mathcal{X}$, en un conjunto de entrenamiento y en un conjunto de prueba. La proporción ha sido 80 % para entrenamiento y el 20 % para pruebas. Posteriormente, cada uno de los conjuntos de entrenamiento han sido agrupados en 7 grupos, aplicando el algoritmo de k-Medias. El resultado de agrupamiento se ha evaluado comparándolo con el agrupamiento representado por las etiquetas del análisis empírico para 7 tipos de fallas, con la finalidad de buscar la semejanza entre ambos agrupamientos. El agrupamiento con mejor evaluación fue el generado por X_4, ya que demostró que 99 % de precisión. Este agrupamiento se seleccionó para descomponerlo en dos partes. La primera parte está compuesta por los datos que representa un alto valor de componente simétrica de secuencia cero, y la segunda parte por las que su valor fue bajo. De la descomposición resulto un agrupamiento de 11 grupos, el cual, fue comparado con el agrupamiento representado por las etiquetas del análisis empírico para 11 tipos de fallas, resultado en el 97.75 % de coincidencia entre ambos agrupamientos.

Capítulo 5

Validación de los modelos obtenidos

En el capítulo anterior se utilizó aprendizaje semi-supervisado [Géron19] para etiquetar al conjunto de entrenamiento de cada conjunto en $\mathcal{X}$. Se han obtenido dos conjuntos de etiquetas, unas para 7 grupos y otro para 11 grupos. En este capítulo se evaluarán los modelos de agrupamiento antes obtenidos mediante el uso de clasificadores. La precisión de los clasificadores nos indicará la calidad del agrupamiento para cada una de las etapas de transformación de datos. En la evaluación usaremos clasificadores con diferentes paradigmas de clasificación. Usaremos vecinos cercanos (k-NN), en el cual no hay entrenamiento, el trabajo principal se invirtió en la fase de la formación de los clusters que en este caso fue k-Medias. Sin embargo, es costoso evaluar la pertenencia de un dato de prueba si el conjunto de datos es muy grande. En la segunda clase de clasificadores usaremos máquinas de soporte vectoriales (tanto lineales (SVM) como su representante no lineal RBF (SVM-RBF)) en la cual la fase de entrenamiento genera una un hiperplano separante en el caso lineal y una hiper-superficie separante en el caso no lineal. Una vez realizado el entrenamiento, la evaluación de pertenencia de un elemento del conjunto de prueba a una de las clases es trivial, ya que solamente tenemos que evaluar si el elemento de prueba está arriba o abajo de la frontera de decisión (en el caso de clasificación binaria). Después usaremos el paradigma de árboles de decisión en el cual el conjunto de discriminantes representados

por cada nodo de los árboles representa la frontera de decisión. La fase de entrenamiento es computacionalmente cara, ya que en esa fase se construye el árbol de decisión, pero la fase de prueba es computacionalmente barata, ya que el elemento sujeto de prueba solo tiene que viajar desde la raíz hasta una de las hojas para determinar la clase a la que pertenece dicho elemento. Adicionalmente, con una ampliación del paradigma de los árboles de decisión, utilizaremos los bosques aleatorios que utiliza la técnica de Bagging para mejorar la eficiencia del clasificador. Finalmente, se utilizará AdaBoost basado en la técnica de boosting para incrementar la eficiencia. Recordemos que la eficiencia de los clasificadores nos indicara también la calidad del agrupamiento que se derivó en este trabajo.

5.1. Clasificadores de datos con 7 etiquetas

Como se mencionó anteriormente, los clasificadores implementados para evaluar los conjuntos en $\mathcal{X}$ son los siguientes: kNN, SVM Lineal, SVM RBF, Árbol de Decisión, Bosque Aleatorio y AdaBoost. Cada uno de los clasificadores son entrenados usando como conjunto de entrenamiento al agrupamiento obtenidos en el capítulo anterior, y después, han sido evaluados usando las métricas de evaluación externa, precisión, exactitud y F1.

5.1.1. Clasificación con datos de $\boldsymbol{X}_1$

Se implementan los clasificadores usando el conjunto de entrenamiento X_1^E y de prueba X_1^P. En la Tabla 5.1 se muestra el resultado obtenido. Los clasificadores que mejor han funcionado son: Árbol de decisión, SVM Lineal y SMV RBF con una clasificación correcta del 40 % de los datos en el conjunto de prueba.

Tabla 5.1: Comparativa entre clasificadores usando X_1

	Entrenamiento				Prueba			
Clasificador	**Score**	**Precision**	**Recall**	**F1**	**Score**	**Precision**	**Recall**	**F1**
kNN	77.53 %	0.52	0.39	0.41	40.0 %	0.26	0.21	0.18
SVM Lineal	100.0 %	1.0	1.0	1.0	40.0 %	0.38	0.26	0.24
SVM RBF	100.0 %	1.0	1.0	1.0	40.0 %	0.33	0.25	0.22
Árbol de decisión	100.0 %	1.0	1.0	1.0	40.0 %	0.27	0.27	0.26
Bosque aleatorio	98.88 %	1.0	0.96	0.98	40.0 %	0.38	0.26	0.24
AdaBoost	79.78 %	0.52	0.56	0.54	40.0 %	0.19	0.21	0.17

5.1.2. Clasificación con datos de X_2

Se implementan los clasificadores usando el conjunto de entrenamiento X_2^E y de prueba X_2^P. En la Tabla 5.2 se muestra el resultado obtenido. El clasificador que mejor ha funcionado es el árbol de decisión con una clasificación correcta del 40 % de los datos en el conjunto de prueba.

Tabla 5.2: Comparativa entre clasificadores usando X_2

	Entrenamiento				Prueba			
Clasificador	**Score**	**Precision**	**Recall**	**F1**	**Score**	**Precision**	**Recall**	**F1**
kNN	77.53 %	0.62	0.49	0.51	25.0 %	0.24	0.29	0.19
SVM Lineal	100.0 %	1.0	1.0	1.0	30.0 %	0.27	0.36	0.24
SVM RBF	100.0 %	1.0	1.0	1.0	20.0 %	0.17	0.18	0.13
Árbol de decisión	100.0 %	1.0	1.0	1.0	40.0 %	0.35	0.46	0.37
Bosque aleatorio	100.0 %	1.0	1.0	1.0	30.0 %	0.31	0.36	0.27
AdaBoost	65.17 %	0.38	0.46	0.41	30.0 %	0.17	0.21	0.15

5.1.3. Clasificación con datos de X_3

Se implementan los clasificadores usando el conjunto de entrenamiento X_3^E y de prueba X_3^P. En la Tabla 5.3 se muestra el resultado obtenido. Los clasificadores que mejor han funcionado son: kNN y SVM Lineal con una clasificación correcta del 60 % de los datos en el conjunto de prueba.

Tabla 5.3: Comparativa entre clasificadores usando X_3

	Entrenamiento				Prueba			
Clasificador	**Score**	**Precision**	**Recall**	**F1**	**Score**	**Precision**	**Recall**	**F1**
kNN	95.51 %	0.96	0.95	0.95	60.0 %	0.38	0.46	0.37
SVM Lineal	98.88 %	0.99	0.99	0.99	60.0 %	0.38	0.46	0.37
SVM RBF	100.0 %	1.0	1.0	1.0	50.0 %	0.36	0.4	0.33
Árbol de decisión	100.0 %	1.0	1.0	1.0	45.0 %	0.34	0.39	0.31
Bosque aleatorio	100.0 %	1.0	1.0	1.0	55.0 %	0.37	0.44	0.35
AdaBoost	60.67 %	0.48	0.57	0.5	35.0 %	0.3	0.27	0.22

5.1.4. Clasificación con datos de X_4

Se implementan los clasificadores usando el conjunto de entrenamiento X_4^E y de prueba X_4^P. En la Tabla 5.4 se muestra el resultado obtenido. El clasificador que mejor ha funcionado es el SVM RBF con una clasificación correcta del 90 % de los datos en el conjunto de prueba.

Tabla 5.4: Comparativa entre clasificadores usando X_7

	Entrenamiento				Prueba			
Clasificador	**Score**	**Precision**	**Recall**	**F1**	**Score**	**Precision**	**Recall**	**F1**
kNN	97.75 %	0.97	0.9	0.92	85.0 %	0.65	0.71	0.68
SVM Lineal	92.13 %	0.64	0.71	0.67	75.0 %	0.51	0.57	0.54
SVM RBF	100.0 %	1.0	1.0	1.0	90.0 %	0.81	0.86	0.83
Árbol de decisión	100.0 %	1.0	1.0	1.0	85.0 %	0.65	0.71	0.68
Bosque aleatorio	100.0 %	1.0	1.0	1.0	85.0 %	0.65	0.71	0.68
AdaBoost	87.64 %	0.48	0.57	0.5	70.0 %	0.36	0.43	0.39

5.2. Clasificadores de datos con 11 etiquetas

5.3. Clasificación con datos de X_5

Los datos X_5 están formados por X_4 agregando una característica que está en función de la componente se secuencia cero. Se implementan los clasificadores para 11 tipos de falla usando el conjunto de entrenamiento X_5^E y de prueba X_5^P. En la Tabla 5.5 se

muestra el resultado obtenido. El clasificador que mejor ha funcionado es SMV RBF con una clasificación correcta del 80 % de los datos en el conjunto de prueba.

Tabla 5.5: Comparativa entre clasificadores usando X_5

	Entrenamiento				Prueba			
Clasificador	**Score**	**Precision**	**Recall**	**F1**	**Score**	**Precision**	**Recall**	**F1**
kNN	91.01 %	0.54	0.57	0.53	80.0 %	0.57	0.62	0.59
SVM Lineal	89.89 %	0.46	0.55	0.49	80.0 %	0.57	0.62	0.59
SVM RBF	100.0 %	1.0	1.0	1.0	80.0 %	0.72	0.73	0.72
Árbol de decisión	100.0 %	1.0	1.0	1.0	75.0 %	0.6	0.6	0.6
Bosque aleatorio	100.0 %	1.0	1.0	1.0	75.0 %	0.6	0.6	0.6
AdaBoost	87.64 %	0.49	0.55	0.5	75.0 %	0.5	0.54	0.51

5.4. Clasificación con datos de X_4 con pos-procesamiento

La clasificación a 11 tipos de fallas es implementada con el conjunto X_4, usando el etiquetado resultante del modelo que incluye las componentes de secuencia cero para el conjunto de entrenamiento X_4^E y el conjunto de pruebas X_4^P. En la Tabla 5.6 se muestra el resultado obtenido. El clasificador que mejor ha funcionado es: SVM RBF con una clasificación correcta del 98 % para los datos en el conjunto de entrenamiento y un 90 % para los datos de prueba.

Tabla 5.6: Comparativa entre clasificadores usando X_4 con pos-procesamiento

	Entrenamiento				Prueba			
Clasificador	**Score**	**Precision**	**Recall**	**F1**	**Score**	**Precision**	**Recall**	**F1**
kNN	95.51 %	0.96	0.96	0.96	85.0 %	0.85	0.85	0.85
SVM Lineal	89.89 %	0.9	0.9	0.9	75.0 %	0.75	0.75	0.75
SVM RBF	97.75 %	0.98	0.98	0.98	90.0 %	0.9	0.9	0.9
Árbol de decisión	97.75 %	0.98	0.98	0.98	85.0 %	0.85	0.85	0.85
Bosque aleatorio	97.75 %	0.98	0.98	0.98	85.0 %	0.85	0.85	0.85
AdaBoost	85.39 %	0.85	0.85	0.85	70.0 %	0.7	0.7	0.7

La clasificación para 11 tipos de falla ha dado mejor resultado al separar fallas bifásicas y trifásicas en aterrizadas y no aterrizadas usando el valor de la componente cero como pos-procesamiento.

5.5. Sistema clasificador de eventos de falla

El análisis, procesamiento y evaluación de las mediciones de un conjunto de eventos ha derivado en una propuesta para etiquetar eventos por tipo de falla a partir de eventos no etiquetado. Estos eventos etiquetados han sido procesados para convertirse en un conjunto de datos útiles para entrenar clasificadores de ML que tengan un buen rendimiento frente a eventos no clasificados.

La propuesta final que ha demostrado tener un mayor éxito de clasificación se presente en forma de esquema en la Figura 5.1.

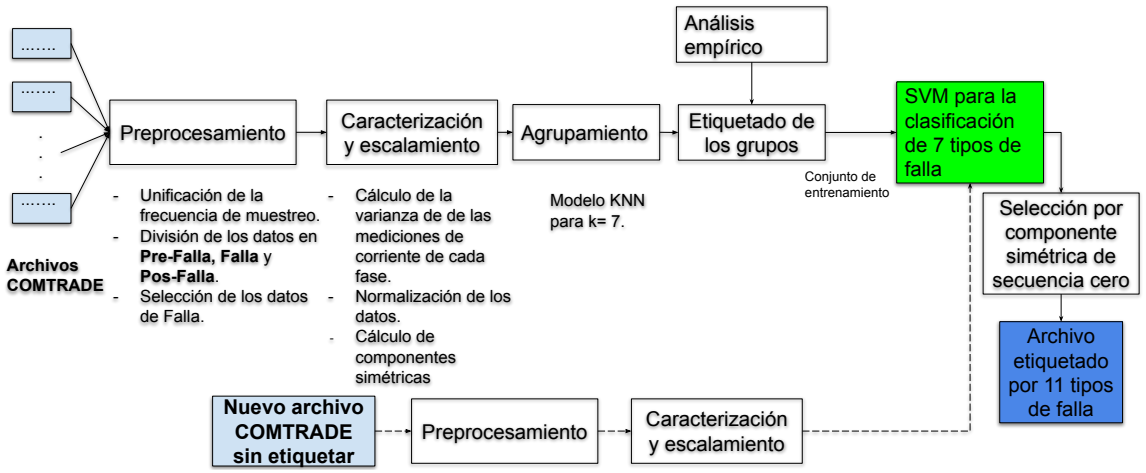


Figura 5.1: Sistema de clasificación de 11 tipos de fallas

Para etiquetar el conjunto eventos se ha propuesto el siguiente proceso:

1. Procesar las mediciones almacenadas en los archivos en formato COMTRADE siguiendo los siguientes pasos:

   (a) Obtener las mediciones en los archivos .dat.

   (b) Unificar su tasa de muestreo a 240 Hz.

   (c) Identificar el inicio de la falla.

   (d) Extraer las mediciones de corrientes pertenecientes a la falla.

   (e) Reconstruir datos de mediciones perdidas por la activación de los sistemas de protección

2. Caracterizar las señales de corriente usando la varianza.

3. Normalizar las mediciones e integrarlos como un conjunto de datos.

4. Agrupar con k-Medias con un valor de k=7

5. Etiquetar grupos por 7 tipos de falla realizando considerando lo siguiente:

   (a) Comparando con un análisis empírico para 7 tipos de fallas

   (b) Dividir por falla aterrizada y no aterrizada utilizando cálculo de componente simétrica de secuencia cero.

El conjunto de datos etiquetados que han sido obtenidos en el proceso anterior será utilizado para entrenar un clasificador SVM RBF. El cual será utilizado para clasificar eventos de falla que tengan el mismo procesamiento, caracterización y normalización que ha llevado al conjunto de entrenamiento.

Capítulo 6

Conclusiones y Trabajo Futuro

El objetivo planteado del presente trabajo es implementar un sistema basado en ML para la identificación y etiquetado de eventos de falla de líneas de transmisión eléctrica por tipo de falla, de manera que sea posible la clasificación de las mismas.

Es necesario mencionar que los datos utilizados para realizar el proceso de clasificación son reales y proviene del registro de eventos de falla ocurridos en líneas de trasmisión eléctricas, de los cuales, se desconoce el tipo de falla a la que pertenece.

Para clasificar los eventos almacenados por tipo de falla usando un clasificador de ML, es necesario construir un conjunto de datos que esté etiquetado por tipo de falla. El conjunto de datos se construyó inicialmente a partir de las mediciones de corriente y voltaje tomadas durante un evento de falla, que están almacenadas en los archivos .dat del formato COMTRADE.

Usando la evaluación de la tendencia al agrupamiento iVAT, se determinó que el conjunto de datos inicial no tenía una buena tendencia, por lo que se reconstruyeron los datos correspondientes a las mediciones posteriores a TRIP, se caracterizaron y se normalizaron para mejorar la evaluación de la tendencia del agrupamiento.

Por otro lado, del conjunto de datos obtenido en el proceso anterior, se derivaron 7 grupos usando el algoritmo de k-Medias con la finalidad de relacionar cada uno de los grupos con un tipo de falla. Lo anterior, permitió etiquetar el conjunto de datos en 3 fallas monofásicas, 3 fallas bifásicas y 1 falla trifásica. De los dos últimos se necesitó determinar

si estaban o no aterrizados. Por lo anterior, se recurre al cálculo de la componente simétrica de secuencia cero, donde se seleccionaron como aterrizada a todas aquellas fallas que la presentaban, resultando en 11 grupos que representan a los 11 tipos de falla. Finalmente, se usó el conjunto de prueba etiquetado y se usó para entrenar a algunos clasificadores teniendo un buen desempeño para la clasificación de eventos.

El enfoque utilizado mediante una estrategia incremental para hacer robustos los datos finalmente resulto en un modelo simple con solamente 4 características, 3 de la cuales son usadas por sistema basado en ML y la otra sirve como discriminaste para detectar si la falla es aterrizada o no lo es, lo cual lo hace robusto

El hecho de que sea simple es también un buen resultado, ya que en algunos trabajos mencionados previamente, se basan en trasformaciones complejas tales como Wavelet, Fourier, etc. Esto es una buena característica del sistema aquí presentado.

Podemos concluir del presente trabajo que se obtuvo un nivel de clasificación aceptable a pesar del pequeño conjunto de datos utilizados, seguramente esto se podrá incrementar en la medida que se tenga acceso a más datos, pero esto depende de los potenciales usuarios de la aplicación. Se estima que con un conjunto de datos ampliado la segunda fase del modelo final quiza no será necesaria, porque el modelo basado en ML podría ahora si realizar mejor la tarea de clasificación.

6.1. Trabajo Futuro

Este trabajo podrá ser utilizado como prueba de concepto ante el potencial usuario, para que de esta forma se ingresen más datos y robustecer este sistema. Una vez hecho esto se obtendrán clases balanceadas. Es necesario continuar con el análisis del sistema de etiquetado y clasificación eventos por tipo de falla, implementado en el presente trabajo, para validar su comportamiento frente a bases de datos más grandes.

Por otro lado, este trabajo puede ser tomado en consideración para solicitar más archivos de fallas para ser procesados y tener un clasificador más robusto que permita ampliar más la certeza de la identificación de fallas en líneas de trasmisión eléctrica utilizando el método propuesto.

Referencias

[A. Asadi Majd17] A. Asadi Majd, H. S. y Ghanbari, T. k-nn based fault detection and classification methods for power transmission systems. *s, Prot. Control Mod. Power Syst*, 2, 2017.

[Abdelgayed18] Abdelgayed, T. S., Morsi, W. G., y Sidhu, T. S. Fault detection and classification based on co-trainingofsemisupervisedmachine learning. *IEEE Transactions on Industrial Electronics*, 65(2):1595–1605, 2018.

[Alvarez09] Alvarez, A. F. *Líneas Eléctricas y trasporte de energía eléctrica*. Universidad Politécnica de Valéncia, 2009. ISBN 978-84-8363-436-3.

[Aritra Dasgupta15] Aritra Dasgupta, a. A. D., Sudipta Debnath. Transmission line fault detection and classification using cross-correlation and k-nearest neighbor. *Engergy System*, (19):183–189, 2015.

[CENACE20] CENACE. *Programa de ampliación y modernización de la red nacional de trasmisión y redes generales de distribución del mercado eléctrico mayorista*. CENACE, 2020.

[Costa14] Costa, F. Fault-induced transient detection based on real-time analysis of the wavelet coefficient energy. *IEEE Trans. Power Deliv*, (140-153):29, 2014.

[Ester96] Ester, M., Kriegel, H.-P., Sander, J., y Xu, X. A density-based

algorithm for discovering clusters in large spatial databases with noise. *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, 96:226–231, 1996.

[Fortescue18] Fortescue, C. L. Method of symmetrical co-ordinates applied to the solution of polyphase networks. *AIEE Transactions*, 37(2):1027–1140, 1918.

[Gangwar20] Gangwar, A. K., Rathore, B., y Mahela, O. P. K-means clustering and linear regression based protection scheme for transmission line. *2020 IEEE 9th Power India International Conference (PIICON)*, págs. 1–6, 2020.

[Géron19] Géron, A. *Hands-on Machine Learning with Scikit-Learn, Keras TensorFlow*. O'Relly Media Inc., 2019. ISBN 978-1-492-03264-9.

[Ikram Sumaiya Thaseen17] Ikram Sumaiya Thaseen, C. A. K. Intrusion detection model using fusion of chi-square feature selection and multi class svm. *Journal of King Saud University - Computer and Information Sciences*, 29(4):462–472, 2017.

[Irwin97] Irwin, J. D. *Análisis básico de circuitos en ingeniería*. Pearson Educación, 1997. ISBN 968-880-816-4.

[J.C. Bezdek22] J.C. Bezdek, R. H. Vat: A tool for visual assessment of cluster tendency. *Proceedings of the 2002 International Joint Conference on Neural Networks*, 185:2225–2230, 2022.

[Jiang11] Jiang, C. C. W. Y., J.A. A hybrid framework for fault detection, classification, and location-part i: concept, structure, and methodology. *IEEE Trans. Power Deliv*, 26(1988-1998), 2011.

[John J. Grainger01] John J. Grainger, W. D. S. J. *Análisis de sistema de potencia*. Graw-Hill, 2001. ISBN 970-10-0908-8.

[Kumar20] Kumar, D. y Bezdek, J. C. Visual approaches for exploratory data analysis. *IEEE SYSTEMS, MAN, CYBERNETICS MAGAZINE*, págs. 10–48, 2020.

[Leonerd90] Leonerd, K. y Peter, R. *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley Sons, 1990.

[Mujal14] Mujal, R. M. *Protección de sistemas eléctricos de potencia*. P, 2014. ISBN 978-84-7653-973-6.

[Ohm27] Ohm, G. S. Die galvanische kette, mathematisch bearbeitet. *Annalen der Physik und Chemie*, 2(8):1–39, 1827.

[Raschka15] Raschka, S. *Python Machine Learning*. Packt Publishing Ltd., 2015. ISBN 978-1-78355-513-0.

[R.N. Mahanty07] R.N. Mahanty, P. D. G. A fuzzy logic based fault classification approach using current samples only. *Electric Power Systems Research*, 77(6-7):501–507, 2007.

[Stevenson85] Stevenson, W. D. *Análisis de Sistemas Eléctricos de Potencia*. McGraw-Hill, 1985. ISBN 0-07-061287-4.

[Tîrnovan19] Tîrnovan, R.-A. y Cristea, M. Advanced techniques for fault detection and classification in electrical power transmission systems: An overview. *8th International Conference on Modern Power Systems (MPS)*, págs. 1–10, 2019.

[Wildi07] Wildi, T. *Máquinas Eléctricas y Sistemas de Potencia*. Pearson Educación, 2007. ISBN 970-26-0814-7.

[Xi23] Xi, Y., Zhang, W., Zhou, F., Tang, X., Li, Z., Zeng, X., y Zhang, P. Transmission line fault detection and classification based on sa-mobilenetv3. *Energy Reports*, 9(2):955–4847, 2023.

[Yang97] Yang, Y. y Pedersen, J. A comparative study on feature selection in text categorization. *Morgan Kaufmann Publishers Inc*, págs. 412–420, 1997.