**Reporte de Avance**

**Información general**

Nombre del proyecto:

**Análisis de Clustering en Series Temporales para la Predicción de Generación Energética de Centrales Solares y Eólicas**

**Fecha del Reporte:** 1 de Marzo de 2024

**Autor:** Ismael Rodríguez Silva

**Objetivos del Proyecto**

Descubrir perfiles o patrones en los datos por central de series temporales que representen fielmente el conjunto de datos de la generación de energía solar y eólica del año 2023.

**Avances y Logros**

**Tareas Completadas**

* **Recopilación de datos**
* **Análisis exploratorio**
* **Transformación y normalización de los datos**
* **Estudio de la literatura**

**En progreso**

* **Algoritmo de clustering (Aplicación de Clustering Fuzzy C-Means)**
* **Medidas de similitud**

**Por hacer**

* **Validación de Clusters**
* **Resultados Preliminares**
  + **Análisis Exploratorio de Datos**
  + **Resultados del Clustering FCM**
  + **Interpretación de los Clusters**

1. **Recopilación de datos**

En primer lugar, descargue y todos los archivos de generación eléctrica solar y eólica del año 2023. Cada mes tiene 6 archivos, con las respectivas series de generación real, pronósticos y vertimiento.

Adicionalmente descargue un archivo con la información de las centrales para obtener la generación potencial máxima bruta, para el posterior cálculo del error.

Mediante programación en Python generé un código que lee todos los archivos e imprime gráficos para visualizar los datos faltantes. Y finalmente genera 2 archivos Excel con los datos consolidados de la energía solar y eólica respectivamente.

En la Tabla 1 se puede visualizar como está estructurado el archivo:

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Real** | **Nombre** | **Fecha** | **Externo** | **Coordinado** | **SExperto** | **Vertimiento** | **Tecnología** | **Región** | **Comuna** | **Potencia\_max\_bruta** |
| 0 | PFV ALMEYDA | 2023-01-01 01:00:00 | 0 | 0 | 0 | 0 | Solar | ATACAMA | Chañaral | 56,8 |
| 0 | PFV ALMEYDA | 2023-01-01 02:00:00 | 0 | 0 | 0 | 0 | Solar | ATACAMA | Chañaral | 56,8 |
| 0 | PFV ALMEYDA | 2023-01-01 03:00:00 | 0 | 0 | 0 | 0 | Solar | ATACAMA | Chañaral | 56,8 |
| 0 | PFV ALMEYDA | 2023-01-01 04:00:00 | 0 | 0 | 0 | 0 | Solar | ATACAMA | Chañaral | 56,8 |
| 0 | PFV ALMEYDA | 2023-01-01 05:00:00 | 0 | 0 | 0 | 0 | Solar | ATACAMA | Chañaral | 56,8 |
| 0 | PFV ALMEYDA | 2023-01-01 06:00:00 | 0 | 0 | 0 | 0 | Solar | ATACAMA | Chañaral | 56,8 |
| 0 | PFV ALMEYDA | 2023-01-01 07:00:00 | 7,9 | 9,4 | 8,596 | 0 | Solar | ATACAMA | Chañaral | 56,8 |
| 0 | PFV ALMEYDA | 2023-01-01 22:00:00 | 0 | 0 | 0 | 0 | Solar | ATACAMA | Chañaral | 56,8 |

Tabla 1 (Elaboración propia)

1. **Análisis exploratorio de datos (AED)**

El AED consiste en examinar los datos para entenderlos, ver relaciones entre variables y encontrar posibles inconsistencias. Para esto, se analiza estadística descriptiva y gráﬁcos.

A continuación, mostraremos el trabajo hecho solo para los datos solares para no extender mucho el reporte.

Para analizar el comportamiento de los datos de las variables numéricas, obtenemos la media, desviación estándar, mínimo, máximo y cuartiles. Mientras que para las variables categóricas obtenemos la cantidad de categorías de la variable (unique), la moda (top) y la frecuencia de de la moda (freq).

Tabla

Descripción generada automáticamente

Se analizó también los datos perdidos o faltantes:

Tabla

Descripción generada automáticamente

Graficamos las variables numéricas para ver la distribución de los datos:

Gráfico, Gráfico de dispersión

Descripción generada automáticamente

Podemos notar que no siguen alguna distribución típica y que están todas con un sesgo hacia la derecha.

Generamos también los diagrama de caja para ver la distribución horaria de las variables.

Una captura de pantalla de un celular con texto e imagen

Descripción generada automáticamente con confianza media

Luego estudiamos las longitudes de las series de tiempo de cada central.

Gráfico

Descripción generada automáticamente

Y podemos ver que hay solo 50 centrales con la data completa, por lo que posteriormente hacemos un filtrado para quedarnos con las centrales con las series temporales completas.

Estudié si las series tenían outliers (valores atípicos) y no encontré.

1. **Transformación y normalización de datos**

Filtre las horas en que las centrales no generan energía, para que cuando se calcule el error estas no lo sobreestimen.

Hice una copia de la data e imputé los valores faltantes con una estrategia de interpolación, para luego correr los modelos con la data con y sin imputación de valores y comparar los desempeños.

Calcule el error con la data sin imputar. A continuación, podemos ver la los gráficos de distribución del error y el de caja y bigote.

Gráfico, Histograma

Descripción generada automáticamente

Gráfico, Gráfico de dispersión

Descripción generada automáticamente

Luego agrupe los datos por fecha y hora.

Gráfico, Histograma

Descripción generada automáticamente

Gráfico, Gráfico de barras

Descripción generada automáticamente

Esto fue realizado también para la data eólica.

Luego transforme las variables a logarítmicas para tratar el sesgo a la derecha y normalice las variables para un posterior mejor desempeño de los modelos.

Imagen que contiene Forma

Descripción generada automáticamente

1. **Estudio de literatura relacionada al Clustering.**

Escribí un documento en Notion con la completitud de mis apuntes, le dejare el [enlace](https://www.notion.so/Time-series-Clustering-Paper-cb7eefc06fe944d5ab9f7ffe6efbaf6a?pvs=4) pero para no extenderme le mencionare lo que encontré más relevante.

Existen 3 tipos de clustering:

Diagrama

Descripción generada automáticamente

En el paper confirma que el subsequence no tiene mucho sentido, por lo que se enfoca principalmente en el “**Whole time series clustering**”.

1. Whole time-series clustering:

* Enfoque: Agrupa series temporales completas en función de sus patrones globales a lo largo del tiempo.
* Proceso: Utiliza el conjunto completo de datos de cada serie temporal para calcular la similitud entre ellas, sin descomponerlas en subsecuencias más pequeñas.
* Representación: Conserva la estructura temporal completa de cada serie, sin reducción de dimensiones.
* Algoritmos Comunes: Puede utilizar k-means adaptado a series temporales, DBSCAN con métrica temporal, o métodos basados en laforma de onda como Dynamic Time Warping Clustering.

Dentro de este tipo de clustering hay 3 enfoques: el basado en estructura, el basado en características y el basado en modelos.

Revisando los existentes trabajos en la literatura, está implícito que lo esencial del clustering de series de tiempo tiene 4 componentes:

1. Reducción de dimensionalidad o método de representación

2. Medida de distancia

3. Algoritmo de clustering

4. Definición del prototipo y evaluación

El proceso general de las series de tiempo usa algunos de estos componentes dependiendo del problema. Usualmente, los datos se aproximan mediante un método de representación de forma que quepan en la memoria. Después, un algoritmo de clustering es aplicada a la data usando la medida de distancia. En el proceso de clustering, usualmente un prototipo es requerido para resumir las series de tiempo. Al final los clusters son evaluados usando criterios.

Después encontré el siguiente paper: “Análisis de las Medidas de Similitud en la Agrupación de Series Temporales para el Descubrimiento de Patrones Energéticos en Edificios”, que tiene objetivos similares al proyecto de clustering que estamos realizando, que es: “Descubrir perfiles o patrones en los datos por central de series temporales que representen fielmente el conjunto de datos de la generación de energía solar y eólica del año 2023. “

En el Paper utilizan el modelo “Fuzzy C-means” y prueba con varias medidas de distancia tales como: “Distancia Euclidiana”, “Correlación de Pearson”, “distancia de Mahalanobis” y “Warping temporal dinámco”.

1. **En progreso**

El plan es aplicar una estrategia similar a la del paper con un algoritmo Fuzzy C-means

Y probar con distintas medidas de distancia.