UNIVERSIDAD DE LOS ANDES

Maestría en Inteligencia Analítica de Datos
APRENDIZAJE NO SUPERVISADO
Ciclo 202214

IDENTIFICACIÓN DE CLUSTERS EN RED CELULAR LTE MEDIANTE APRENDIZAJE NO SUPERVISADO (ML)

David Santiago Muñoz Carrillo. Código: 202121443
Andrés Felipe Arteta Isaacs. Código: 201124652
Iván Camilo Barriga Gómez Código: 202121106
Edison Arcángel Giraldo Martínez. Código: 202124624

Tabla de Contenido

Resumen

Un proveedor de servicio de telefonía celular LTE ha recolectado información física de las estaciones radiantes, parámetros de configuración y desempeño de indicadores de calidad (KPI) para orientar las estrategias de planificación y optimización de la red nacional, con el fin de maximizar la calidad del servicio entregado a los usuarios. Con base en estos datos, se aplican diferentes algoritmos de agrupación, como K-medias y K-medoides, para responder a las siguientes preguntas: ¿Podrían detectarse características comunes en las celdas con alta y baja velocidad de navegación? ¿Existe algún tipo de asociación entre celdas de acuerdo con el desempeño de los principales KPI? Como principal resultado se encontró que Kmedias y K-medoides lograron agrupar las celdas en dos clústeres, las de buen desempeño (mayor velocidad descarga y mayor disponibilidad) y las de bajo desempeño. Estos patrones encontrados, brindan al área de negocio información valiosa sobre las características de los grupos de celdas, que generan puntos de acción correctiva y de planeación en las estrategias de optimización de la red para maximizar la calidad del servicio prestado al usuario.

Introducción

La red de telefonía celular está compuesta por múltiples estaciones radiantes, que brindan cobertura en áreas delimitadas conocidas como celdas. La celda permite la comunicación inalámbrica por radiofrecuencia entre los equipos móviles y la estación radiante. Esta red se encuentra en constante cambio, principalmente por los trabajos de expansión de infraestructura, optimización de la red y cambios en la demanda de servicios de voz y datos, por lo que realizar labores de monitoreo y control son fundamentales para garantizar el correcto funcionamiento y aseguramiento de la calidad del servicio por parte del operador.

Un operador de telecomunicaciones privado de Colombia tiene más de 70 mil celdas que soportan la operación en la red LTE. Cada celda brinda el servicio de voz y datos a los usuarios, generando métricas de calidad del servicio. Además, existe un registro con información de identificación y características físicas de la celda y otro registro de los valores de parametrización, los cuales serán la base para realizar un análisis de ML no supervisado que permita explorar e identificar posibles patrones de grupos de celdas con determinados comportamientos. Por ejemplo, identificar si existen grupos de celdas con características similares que brinden mejor velocidad de descarga que otras, o evidenciar si, por el contrario, hay algún grupo de celdas que presentan baja disponibilidad, es decir, que por alguna razón, están teniendo fallas de fuera de servicio y, por último, grupos de celdas con mayor proporción de sesiones exitosas. Todo lo anterior puede brindar información sobre aquellas características “deseadas” de las celdas que permitan prestar un mejor servicio.

El caso de estudio está utilizando datos de 24 horas de un solo día, en agregación diaria, lo que nos proporciona una mirada específica del comportamiento de la red en ese día. Escalar el caso de estudio para analizar más días supone un límite y a su vez varios retos. Por ejemplo, es necesario incluir la temporalidad ya que es posible que algunos parámetros de celdas cambien en el transcurso del tiempo. Existe otro reto importante de capacidad, ya que se aumenta la cantidad de data cruda para

preprocesar de manera proporcional a la cantidad de días a estudiar. Para poner en contexto, la fórmula de conteo de registros sería 70.000 celdas x 24 horas x Cantidad de días, y lo anterior agregado por día.

Debido a la necesidad de procesar información en las múltiples áreas de una red celular, existen diversas aplicaciones con técnicas de machine learning, según el caso de negocio a resolver. Existen algunas publicaciones [1][2] que hablan sobre la importancia de las técnicas de ML para el avance en las tecnologías de telecomunicación, destacando aplicaciones tales como la detección de anomalías, análisis de causa raíz, administración y optimización de la red, mantenimiento predictivo, gestión de incidencias y abandono churn. Hay también diversos trabajos, aplicados tanto a nivel nacional [3], como internacional [4] donde se combinan técnicas de aprendizaje supervisado y no supervisado para explorar y optimizar las características de redes de telecomunicación.

Materiales y Métodos

Descripción de las fuentes de datos

Para capturar la información de la data cruda, se consultaron 2 fuentes privadas que contienen datos estructurados de identificación y de parametrización de todas las celdas de la red nacional y sus KPI para del día 23 de agosto de 2022.

Descripción de los datos

Las siguientes variables corresponden a la identificación y características fijas de la celda: Nombre, Frecuencia de radiación, Modelo de hardware, Altura antena e Inclinación antena.

También se tienen variables que corresponden a las características de parametrización de la celda: Potencia de radiación, Ancho de banda, Máximo número de usuarios, Código de zona y Multiantena.

Finalmente, tenemos las variables correspondientes a los KPI: Velocidad de Navegación, Disponibilidad de la celda y Navegación exitosa.

Preprocesamiento de Datos

Debido a que los registros de interés están ligados a los KPIs de la prestación efectiva del servicio al usuario, los registros vacíos que no brindan servicio no son de interés, y, por lo tanto, se realiza un proceso de limpieza eliminando los registros con datos nulos (NaN).

También hay instancias donde la información de la celda se actualiza más de una vez en el mismo día, por lo que pueden aparecer varios registros de la misma celda (nombre). En este caso, se eliminan los duplicados y se deja el último registro, que corresponde a la información más actualizada de la celda.

Finalmente, se corrige el tipo de dato de algunas columnas que no corresponde a su naturaleza (Potencia de radicación y Máximo número de usuarios), debido principalmente a datos numéricos almacenados como texto.

Para tener los datos de entrada listos para los modelos a aplicar, se hace encoding de las columnas categóricas: One-Hot Encoding para ‘Ancho de banda', y 'Multiantena’, y BinaryEncoder para las demás columnas (para evitar generar más dimensiones). Con este cambio, tenemos una matriz de datos de 70371 filas y 34 columnas. Dado que se generan

columnas con datos en diferente escala, generamos adicionalmente otra tabla con los datos escalados, apoyándonos de la función StandardScaler de Scikit-learn.

Estadísticas básicas de los datos

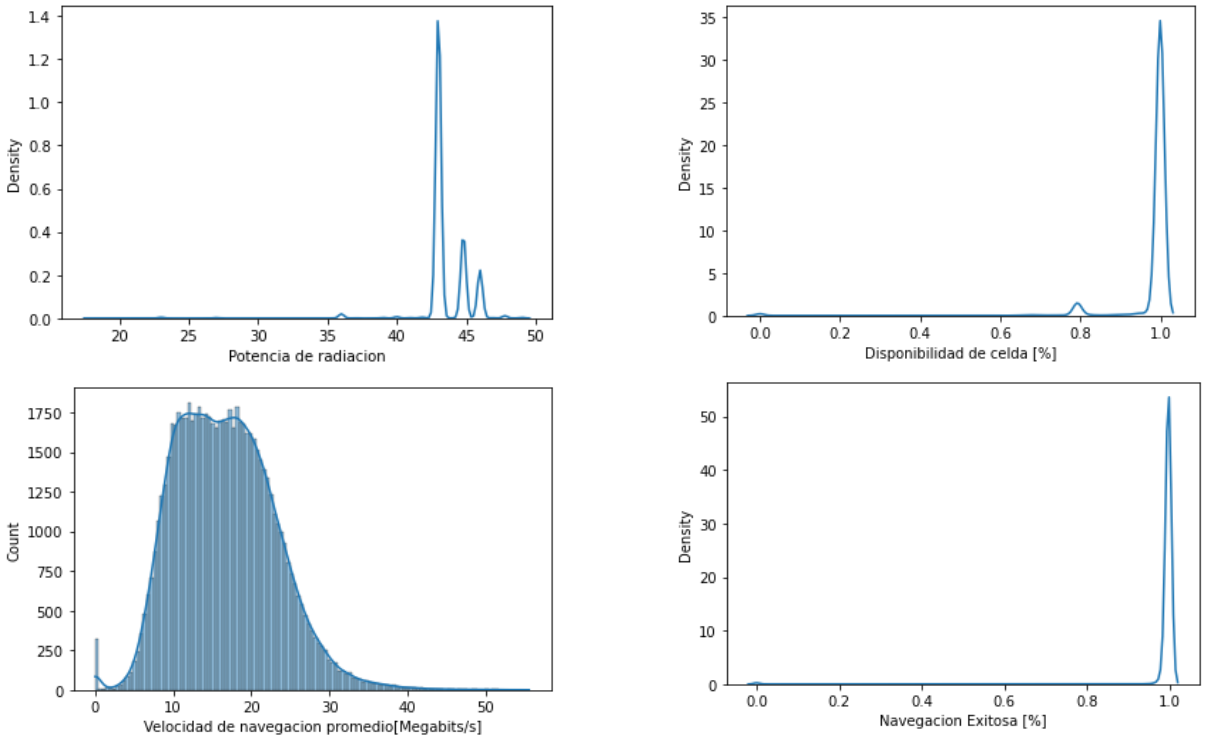
Las estadísticas básicas de los datos se muestran en las tablas 1 y 2. También se muestran algunas gráficas con el comportamiento de algunas variables de interés.

	Media	Desv. est.	min	25%	50%	75%	max
Altura Antena	31,26	14,48	0,00	21,00	30,00	40,00	100,00
Inclinacion Antena	5,72	2,54	0,00	4,00	6,00	7,00	19,00
Potencia de radiación	43,54	1,76	18,00	43,00	43,00	44,00	49,00
Maximo numero de Usuarios	905,34	318,75	230,00	650,00	960,00	1.200,00	1.470,00
Velocidad de navegación promedio	16,64	6,37	0,00	11,78	16,25	20,85	55,56
Disponibilidad de celda	0,9785	0,0970	0,0000	0,9991	1,0000	1,0000	1,0000
Navegacion Exitosa	0,9932	0,0628	0,0000	0,9968	0,9986	0,9992	1,0000

Tabla 1. Estadísticas descriptivas variables numéricas

	No. de valores únicos	Valor más frecuente	Frecuencia
Frecuencia de radiación	5	1900	20.741
Modelo del hardware	167	ASI4517R3v18	10.357
Ancho de banda	4	15 MHz	40.520
Codigo de Zona	252	27012	1.668
Multiantena	6	Closed Loop MIMO (4x4)	34.170

Tabla 2. Estadísticas descriptivas variables categóricas



Gráfica 1. Distribución de variables de interés

Al analizar las tablas y gráficas anteriores, podemos observar varios comportamientos interesantes. En primera medida, una potencia de radiación relativamente uniforme entre las celdas, salvo por algunas en el espectro inferior de la distribución, que irradian muy por debajo de la media.

También vemos que la velocidad de navegación es el KPI que mayor dispersión presenta, con algunas pocas celdas llegando a velocidades muy altas, más de dos veces por arriba de la media.

En cuanto a la disponibilidad de celda, la gran mayoría tienen un valor de 1 (o 100%), aunque hay algunas pocas alrededor de 80% y 0%. De igual forma, el porcentaje de navegación exitosa es cercano a 100% en la mayoría de los casos, salvo algunos pocos en 0%.

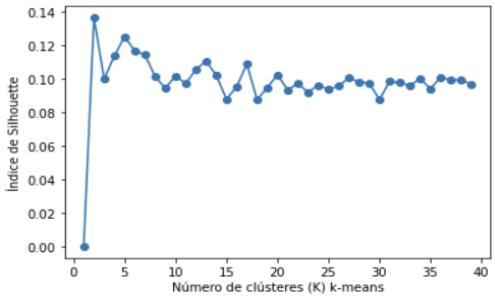
Resultados y Discusión

Análisis de Componentes Principales (PCA)

El PCA fue realizado con el apoyo de la librería Numpy y la función linalg.svd. Al analizar la varianza explicada por los componentes, vemos que con 24 componentes se explica el 90% de la varianza, y para explicar 98% requerimos 29. Por lo tanto, continuamos con la matriz de datos completa para los siguientes modelos, por considerar que no se tiene una ganancia significativa al descomponer la matriz de datos.

K-medias

Para el algoritmo de K-medias nos apoyamos de la librería Scikit-learn y la función KMeans, el cual utiliza la distancia euclidiana para asignar los clústeres. Para elegir la cantidad de clúster iniciales, utilizamos el índice de Silhouette, con el que encontramos el máximo en 2, aunque también hay un valor alto en 5, según se ve en la siguiente figura:



Para efecto de tener una mayor diversidad de clústeres, elegimos k=5. En la tabla 3 se presentan los promedios de las columnas numéricas para cada clúster.

La mayor diferenciación de los grupos se evidencia en la frecuencia de radiación. Podríamos hablar de que algunas características particulares de los clústeres:

- Los clústeres 0 y 1 tienen las frecuencias de radiación más altas, pero el 1 tiene una diferencia apreciable en el número máximo de usuarios. En general, el clúster 1 tiene menores valores para las demás variables.
- El clúster 2 tiene valores medios de frecuencia de radiación, y con la mayor altura de antena y potencia de radiación. Tiene el menor valor de Máximo número de usuarios y tiene la menor velocidad de navegación.
- Los clústeres 3 y 4 tienen valores medio de velocidad de navegación, pero disponibilidad de celda y navegación más bajos que los demás clústeres. El 3 particularmente, destaca que tiene una frecuencia de radiación alta, pero tiene los menores valores de disponibilidad de celda y navegación exitosa.

Clúster	Frecuencia de radiación	Altura Antena	Inclinación Antena	Potencia de radiación	Máximo número de Usuarios	Velocidad de navegación promedio[Megabits/s]	Disponibilidad de celda [%]	Navegación Exitosa [%]
0	2.566,33	31,80	5,43	42,96	1.189,72	21,43	99,27%	99,88%
1	2.476,38	25,45	5,13	44,39	716,88	17,82	99,08%	99,74%
2	854,21	46,02	6,27	45,57	418,80	5,53	99,61%	98,72%
3	1.584,08	33,12	6,02	43,71	695,58	12,16	95,22%	98,49%
4	700,97	34,01	7,12	43,28	951,66	11,29	98,76%	99,11%

Tabla 3. Clústeres obtenidos con k-means

K-medoides

Para el algoritmo de K-medoides nos apoyamos de la librería Scikit-learn y la función KMedoids, y utilizamos la matriz de distancia de Gower, especialmente utilizada cuando hay variables mixtas (cuantitativas y categóricas). Se calcula la matriz cuadrada de gower y se pasa como argumento del .fit al algoritmo Kmedoids. Se establece nuevamente el parámetro clúster=2 y se obtienen las siguientes características de los resultados:

- Se observa que el clúster 0 representa todas aquellas celdas con mayores velocidades de descarga y mayor disponibilidad de la celda. En contraparte, el clúster 1 representa las celdas con menores velocidades de descarga y menor disponibilidad.
- El clúster 0 contiene una mayor cantidad de celdas con frecuencias altas y el clúster 1 con mayor cantidad de frecuencias bajas.
- El clúster 0 contiene mayor proporción de celdas con una configuración de máximo número de usuarios Alta y el clúster 1, configuración de máximo de usuarios baja.

Clúster	Frecuencia de radiación	Altura Antena	Inclinación Antena	Potencia de radiación	Máximo número de Usuarios	Velocidad de navegación promedio[Megabits/s]	Disponibilidad de celda [%]	Navegación Exitosa [%]
0	2.518,54	30,41	5,32	43,30	1.057,83	20,23	98,93%	99,57%
1	1.509,88	32,22	6,18	43,81	730,78	12,53	96,62%	99,03%

Tabla 4. Clústeres obtenidos con k-medoids

Conclusión

• Trabajar con más de 2 dimensiones representa un reto importante especialmente en clustering ya que para realizar la interpretación de los resultados es común hacerlo mediante gráficos de dispersión en dos o tres dimensiones, y en este ejercicio no se logró la reducción a componentes principales.

• En Kmedoides y Kmeans se obtuvieron resultados similares cuando se seleccionaron 2 clúster los cuales permitieron diferenciar el conjunto de celdas debido a las características y kpis de desempeño, en donde las celdas del clúster 0 tienen un mejor desempeño que las celdas del clúster 1.

• Se logró identificar en los 2 grupos de celdas, diferencias notables en ciertas características importantes como la frecuencia de radiación y la configuración máxima de usuarios. Esta información es valiosa para el negocio ya que es posible orientar la investigación y análisis profundos con estos grupos de celdas segmentadas.

Referencias

[1] Haidine, Abdelfatteh et al. "Artificial Intelligence and Machine Learning in 5G and beyond: A Survey and Perspectives". Moving Broadband Mobile Communications Forward - Intelligent Technologies for 5G and Beyond, edited by Abdelfatteh Haidine, IntechOpen, 2021. 10.5772/intechopen.98517. Artificial Intelligence and Machine Learning in 5G and beyond: A Survey and Perspectives | IntechOpen

[2] H. Huang, W. Xia, J. Xiong, J. Yang, G. Zheng and X. Zhu, "Unsupervised Learning-Based Fast Beamforming Design for Downlink MIMO," in IEEE Access, vol. 7, pp. 7599-7605, 2019, doi: 10.1109/ACCESS.2018.2887308. https://ieeexplore.ieee.org/abstract/document/8586870/

[3] Ordoñez, Marco A. “Optimización de redes UMTS soportada en Machine Learning”. 2021. Universidad Distrital Francisco José De Caldas, tesis de Maestría. Recuperado de Optimización de redes UMTS soportada en Machine Learning (udistrital.edu.co)

[4] Nikita Butakov, Loren Jan Wilson, Wenting Sun, Angel Barranco(2021) “Machine learning use cases: how to design ML architectures for today’s telecom systems”. Recuperado de https://www.ericsson.com/en/blog/2021/5/machine-learning-use-cases-in-telecom.