Room Occupancy Stimation

Daniel Alejandro Yepes Mesa, alejandro.yepes@udea.edu.co
Juan Felipe Santa Ospina juan.santa@udea.edu.co
Cristian David Tamayo Espinosa cristian.tamayoe@udea.edu.co
Dpto. of Systems Engineering and Computer Science
Universidad de Antioquia

1. Descripción del Contexto del Problema y Utilidad de la Solución Basada en ML

Estimar con precisión el número de ocupantes en una sala se vuelve imprescindible, ya que el consumo energético ejerce un papel determinante en el gasto global, siendo los sistemas de calefacción, ventilación, aire acondicionado e iluminación sus principales responsables. Por lo tanto, es primordial definir una estrategia efectiva para optimizar el uso de estos sistemas y ajustarlos dinámicamente al número de ocupantes en un espacio, para de este modo además de optimizar el uso de estos, se logre estimar correctamente en función de las necesidades del espacio y así obtener un ahorro energético. Esto se aborda utilizando sensores ambientales no intrusivos, estos sensores ofrecen una alternativa que respeta la privacidad para recopilar datos relevantes sobre el entorno de una sala, y a asi evitar acudir a otros metodos intrusivos como cámaras, que pueden generar preocupaciones con respecto a la privacidad. Por lo anterior, la utilidad de desarrollar una solución de Machine Learning para este problema es multifacética, ya que al ajustar dinámicamente los sistemas según la presencia real se permite una considerable eficiencia energética lo que mejora simultáneamente el confort de los ocupantes, esto a su vez también impulsa la automatización inteligente de edificios y facilita un valioso análisis para la optimización del uso del espacio, permitiendo así estimaciones de ocupación más robustas y precisas.

2. Descripción de la Composición de la Base de Datos

El dataset utilizado para este proyecto es el "Room Occupancy Estimation Dataset" del UCI Machine Learning Repository [1].

Número de Muestras: El dataset contiene 10,129 instancias. Cada instancia representa un conjunto de lecturas de sensores tomadas en un intervalo de 30 segundos durante un período de 4 días.

Número de Variables: Hay un total de 19 variables, de estas, 18 son características (features) y 1 es la variable objetivo (target).

Significado de las Variables:

Date (Fecha): Fecha de la medición (YYYY/MM/DD). Tipo: Fecha.

Time (Hora): Hora de la medición (HH:MM:SS). Tipo: Fecha/Hora.

S1_Temp a S4_Temp: Lecturas de temperatura de 4 sensores distintos (S1 a S4). Tipo: Continuo. Unidades: Grados Celsius (°C).

S1_Light a S4_Light: Lecturas de intensidad lumínica de los mismos 4 sensores. Tipo: Entero. Unidades: Lux.

S1_Sound a S4_Sound: Lecturas del nivel de sonido (salida del amplificador leída por ADC) de los mismos 4 sensores. Tipo: Continuo. Unidades: Volts.

S5_CO2: Nivel de dióxido de carbono medido por el sensor S5. Tipo: Entero. Unidades: Partes por millón (PPM).

S5_CO2_Slope: Mide qué tan rápido está cambiando el nivel de CO2. Ayuda a detectar más rápidamente la presencia de personas. Es un valor numérico continuo.

S6_PIR, S7_PIR: Valores de los sensores infrarrojos pasivos que detectan movimiento. Tipo: Binario (0 o 1, para detección de movimiento). Aunque S7_PIR está listado como entero en UCI su función es binaria.

Room_Occupancy_Count (Variable Objetivo): Número real de ocupantes en la sala. Rango: 0 a 3 personas. Tipo: Entero.

Según la descripción del dataset en UCI, no hay valores faltantes, y dado que no hay datos faltantes, no se requiere una estrategia de imputación en esta etapa.

Codificación de Variables:

Date y Time: Actualmente están como cadenas de texto o tipos de fecha/hora. Para su uso en modelos, podrían transformarse en características numéricas

más útiles como por ejemplo hora del día, día de la semana, etc.

Variables de Sensores (Temp, Light, Sound, CO2, CO2_Slope): Son numéricas, y pueden usarse directamente, aunque la normalización o estandarización será considerada.

Variables PIR (S6_PIR, S7_PIR): Son inherentemente binarias (0 para no movimiento, 1 para movimiento).

Variable Objetivo (Room_Occupancy_Count): Es una variable entera que representa categorías ordenadas (0, 1, 2, 3).

3. Tipo de Configuración o Paradigma de Aprendizaje y su Justificación

Para abordar el problema de estimación de ocupantes en la sala, el equipo ha decidido que el paradigma de aprendizaje más apropiado es el de aprendizaje supervisado imbalance learning, la justificación de esta elección se basa, en que, es aprendizaje supervisado debido a la naturaleza del dataset proporcionado, dado que para para cada conjunto de entradas, conocemos la salida que es el número de ocupantes, por lo que el objetivo es entrenar modelos que relacionen dichas lecturas de los sensores con el número de ocupantes. Por otra parte, imbalanced learning, debido a que al observar el dataset se observa una distribución de clases significativamente sesgada, con una predominancia de la clase correspondiente a 0 ocupantes, por lo que las estrategias para mitigar el impacto de este desbalance serán una consideración fundamental en el desarrollo y evaluación del modelo.

4. Estado del arte

Empezamos analizando el artículo, **Room Occupancy Prediction: Exploring the Power of Machine Learning and Temporal Insights [2]**. Este trabajo propone una solución al problema de estimar la ocupación de habitaciones utilizando diversos modelos de clasificación. Los autores trabajan directamente con el conjunto de datos "Room Occupancy Estimation" del UCI [1] y evalúan una serie de algoritmos de aprendizaje automático para predecir la cantidad de ocupantes en una habitación, a partir de sensores de temperatura, luz, sonido, CO_2 y movimiento. Paradigma de aprendizaje: Supervisado, clasificación multiclase. Las técnicas utilizadas fueron: Regresión logística multinomial, Análisis Discriminante Lineal (LDA), Máquinas de Vectores de Soporte (SVM), Random Forest, XGBoost, LightGBM y Perceptrón Multicapa (MLP). Las metodología de validación fueron: validación cruzada de 5 pliegues junto con ajuste de hiperparámetros utilizando búsqueda en malla. En el apartado de métricas de evaluación tenemos Precisión (accuracy), AUC ponderado y análisis SHAP para interpretación de resultados. El modelo que obtuvo el mejor desempeño fue Random Forest, alcanzando una precisión del 98.7%. Los autores destacan que las variables más relevantes fueron las de iluminación (especialmente S1_Light y S2_Light), lo que sugiere que la intensidad lumínica está estrechamente relacionada con la presencia de personas. Además, el análisis SHAP permitió comprender cómo cada variable contribuía a la predicción, facilitando la interpretación del modelo y resaltando la importancia de ciertos sensores ambientales. A pesar de no modelar explícitamente la temporalidad, se observó que la alta frecuencia de muestreo de los datos capturaba implícitamente patrones temporales, mejorando la capacidad predictiva de modelos tradicionalmente no secuenciales.

Continuamos con **Ubiquitous Multi-Occupant Detection in Smart Environments [3].** Este estudio se enfoca en la detección de múltiples ocupantes en entornos inteligentes utilizando sensores no invasivos y modelos de aprendizaje profundo que capturan la dinámica temporal de las señales. Aunque no se menciona explícitamente que se utilizó la misma base de datos del UCI [1], la estructura de datos, tipo de sensores y objetivo del estudio son altamente similares. El paradigma de aprendizaje predominante es Supervisado, con enfoque secuencial (serie de tiempo). Las técnicas utilizadas fueron los modelos tradicionales como SVM, QDA y k-NN, y modelos de aprendizaje profundo como MLP, LSTM, GRU y BiGRU. Ahora, las metodologías de validación utilizadas son las de separación entre conjuntos de entrenamiento y prueba mediante ventanas deslizantes, permitiendo evaluar el rendimiento en secuencias temporales.Ahora para evaluar el modelo usaron métricas como: Precisión, F1-score, recall y exactitud. Las métricas F1-score y recall permiten evaluar el rendimiento de clasificación considerando el balance entre clases. El modelo BiGRU fue el más destacado, obteniendo un F1-score superior al 90% en la tarea de estimar múltiples ocupantes. Este resultado demuestra la ventaja de utilizar modelos secuenciales en problemas donde los datos presentan dependencia temporal. El trabajo resalta que, al modelar la información contextual a través del tiempo, se mejora significativamente la capacidad del sistema para detectar no solo la presencia, sino también el número exacto de personas en una habitación. Los modelos tradicionales mostraron buen desempeño, pero fueron superados por las redes neuronales recurrentes en tareas de mayor complejidad.

Bibliografía

[1] A. Singh and S. Chaudhari. "Room Occupancy Estimation," UCI Machine Learning Repository, 2018. [Online]. Disponible en: https://doi.org/10.24432/C5P605.

[2] S. Mao, Y. Yuan, Y. Li, Z. Wang, Y. Yao y Y. Kang, "Room Occupancy Prediction: Exploring the Power of Machine Learning and Temporal Insights," *American Journal of Applied Mathematics and Statistics*, vol. 12, no. 1, pp. 1–9, 2024. [En línea]. Disponible en: https://pubs.sciepub.com/ajams/12/1/1/index.html

[3] D. Fährmann, F. Boutros, P. Kubon, F. Kirchbuchner, A. Kuijper y N. Damer, "Ubiquitous Multi-Occupant Detection in Smart Environments," *Neural Computing and Applications*, vol. 36, no. 6, pp. 2941–2960, 2024. [En línea]. Disponible en: https://link.springer.com/article/10.1007/s00521-023-09162-z