UNIVERSIDAD NACIONAL DE SAN ANTONIO ABAD DEL CUSCO
FACULTAD DE INGENIERÍA ELÉCTRICA, ELECTRÓNICA, INFORMÁTICA Y MECÁNICA
ESCUELA PROFESIONAL DE INGENIERÍA INFORMÁTICA Y DE SISTEMAS

TESIS

"ANÁLISIS COMPARATIVO DE LA PERFORMANCE DE LOS DESCRIPTORES WAVELET Y FOURIER, APLICADO A LA DETECCIÓN DE ANOMALÍAS EN TRAYECTORIAS"

Para optar al título profesional de:
INGENIERO INFORMÁTICO Y DE SISTEMAS

Presentado por:
Br. GERAR FRANCIS QUISPE TORRES

Asesor:
Dr. LAURO ENCISO RODAS

Financiado: Convenio Marco
CONCYTEC-UNSAAC-FONDECYT

Cusco - Perú,
2022

Acrónimos

AP Affinity Propagation

AUC Area Under the Curve

CPD Closest Pair Distance

CLEI Centro Latinoamericano de Estudios de Informáctica

CLEIej CLEI electronic journal

DBSCAN Density-based spatial clustering of applications with noise

DFT Discrete Fourier Transform

DTW Dynamic Time Warping

ED Euclidean Distance

FPR False Positive Rate

GPS Global Positioning System

GPU Graphics Processing Unit

ITS Intelligent Transportation System

kNN K-Nearest Neighbor

LCSS Longest Common Sub-Sequence

MDWD Multilevel Discrete Wavelet Decomposition

PCA Principal component analysis

ROC Receiver Operating Characteristic

STI Sistema de Transporte Inteligente

SPD Sum of Pair Distance

t-SNE t-Distributed Stochastic Neighbor Embedding

TPR True Positive Rate

Agradecimientos

Deseo agradecer a todos los que me ayudaron con este estudio.

A mi asesor y co-asesores, quienes me ayudaron con sus conocimientos y consejos; al *Laboratorio de Algoritmos y Análisis de Datos* (LAAD) de la Universidad Nacional de San Antonio Abad del Cusco - UNSAAC, cuyo personal de investigación me apoyó para así terminar el presente trabajo.

También quisiera agradecer a mi familia, a mi madre y mis hermanos por sus apoyos incondicionales.

Resumen

El procesamiento automático de trayectorias tiene diversas aplicaciones en campos como la meteorología, el tráfico marítimo, la seguridad y el seguimiento de objetos. En este estudio, se aborda el problema de la detección de trayectorias anómalas mediante el análisis de su morfología. El objetivo es comparar dos descriptores de trayectorias basados en las transformadas Wavelet y Fourier en términos de su capacidad para detectar anomalías. La justificación de este estudio se basa en la necesidad de explorar a fondo el uso de descriptores de trayectorias para la detección de anomalías. Aunque existen investigaciones previas en este campo, aún no se ha realizado un análisis exhaustivo de la detección de anomalías basada en la morfología de las trayectorias. La metodología utilizada en este estudio se basa en un enfoque cuantitativo. Se recopilaron y utilizaron bases de datos presentes en la literatura, incluyendo la base de datos creada por Piciarelli, la base de datos CROSS y dos bases de datos nuevas. Se aplicaron las transformadas Wavelet y Fourier para obtener descriptores de trayectorias, y se utilizó un método de aprendizaje no supervisado para el agrupamiento de trayectorias. Los resultados obtenidos revelaron que el descriptor de trayectorias basado en las transformadas Wavelet mostró un mejor rendimiento en la detección de anomalías en comparación con el descriptor basado en la transformada de Fourier. Se observó una mayor capacidad para discriminar y detectar trayectorias anómalas al considerar la morfología de las mismas mediante las transformadas Wavelet. En conclusión, este estudio contribuye al campo de la detección de anomalías al explorar y comparar descriptores de trayectorias basados en las transformadas Wavelet y Fourier. Se destaca la importancia de considerar la morfología de las trayectorias para una detección más precisa. El enfoque cuantitativo utilizado en este estudio proporciona una base sólida para futuras investigaciones en el campo del procesamiento automático de trayectorias y la detección de anomalías.

Palabras clave: Detección de trayectorias anómalas, análisis comparativo, descriptor de morfología, extracción de características, descriptor de trayectorias.

Abstract

Automatic trajectory processing has various applications in fields such as meteorology, maritime traffic, security, and object tracking. This study addresses the problem of detecting anomalous trajectories through the analysis of their morphology. The objective is to compare two trajectory descriptors based on Wavelet and Fourier transforms in terms of their ability to detect anomalies. The justification for this study is based on the need to thoroughly explore the use of trajectory descriptors for anomaly detection. Although previous research exists in this field, there has not yet been a comprehensive analysis of anomaly detection based on trajectory morphology. The methodology used in this study is based on a quantitative approach. Databases present in the literature were collected and used, including the database created by Piciarelli, the CROSS database, and two new databases. Wavelet and Fourier transforms were applied to obtain trajectory descriptors, and an unsupervised learning method was used for trajectory clustering. The results revealed that the trajectory descriptor based on Wavelet transforms showed better performance in anomaly detection compared to the descriptor based on Fourier transform. A greater ability to discriminate and detect anomalous trajectories was observed by considering the morphology of the trajectories through Wavelet transforms. In conclusion, this study contributes to the field of anomaly detection by exploring and comparing trajectory descriptors based on Wavelet and Fourier transforms. The importance of considering trajectory morphology for more precise detection is emphasized. The quantitative approach used in this study provides a solid foundation for future research in the field of automatic trajectory processing and anomaly detection.

Keywords: Anomalous trajectory detection, comparative analysis, morphology descriptor, feature extraction, trajectory descriptor.

Índice General Páginas

Índice General **XIV**

Índice de Tablas **XV**

Índice de Figuras **XVII**

Índice de Algoritmos **XIX**

Introducción **1**

1. Aspectos Generales **3**

1.1. El problema de Investigación 3

1.1.1. Descripción del Problema 4

1.1.2. Identificación del Problema 5

1.2. Antecedentes 5

1.2.1. Introducción al análisis de trayectorias 5

1.2.2. Procesamiento de trayectorias 6

1.2.3. Detección de anomalías 7

1.2.4. Conjunto de datos utilizados 7

1.3. Objetivos 8

1.3.1. Objectivo General 8

1.3.2. Objetivos Específicos 8

1.4. Hipótesis 8
1.4.1. Definición de Variables 9
1.5. Justificación 9
1.6. Métodos de Evaluación 11
1.7. Alcances y Limitaciones 11
1.7.1. Alcances 11
1.7.2. Limitaciones 12
1.8. Metodología de Investigación 12
1.9. Contribuciones de esta tesis 14
1.9.1. Contribuciones en el Ámbito Académico 14
1.9.2. Contribuciones Generales 14

2. Marco Teórico **15**
2.1. Datos de Trayectoria 15
2.2. Números Complejos 16
2.2.1. Adición y Sustracción de Números complejos 16
2.2.2. Multiplicación y División de Números Complejos 16
2.3. Histogramas 17
2.3.1. Tipos de Histogramas 17
2.4. La Transformada de Fourier 20
2.4.1. La Transformada Discreta de Fourier (DFT) 20
2.5. La transformada de Wavelet 21
2.5.1. La Descomposición Multinivel Discreta de Wavelet 21
2.6. Affinity Propagation (AP) 23
2.7. K vecinos más próximos ó K-Nearest Neighbor (K-NN) 24
2.7.1. Ventajas y Desventajas del Algoritmo k-NN 25
2.8. T-Distributed Stochastic Neighbor Embedding (t-SNE) 26

2.9. Trayectoria Anómala 27
2.10. Medidas de Similaridad 28
2.11. Demarcado (Map Matching) 28
2.12. Longitud de Trayectoria 29
2.13. Tasas de Muestreo 29
2.14. Dirección y Regiones 29
2.15. El termino "performance" dentro de la tesis 29

3. Desarrollo del trabajo de investigación 31
3.1. Datos de trayectorias 31
3.2. Modelamiento de Trayectorias 33
3.2.1. *Normalización de Trayectoria* 33
3.2.2. *Descomposición de Trayectoria* 35
3.2.3. *Representación de Espacio de Características* 35
3.3. Detección de Anomalía 38

4. Pruebas y resultados 39
4.1. Configuración Experimental 39
4.2. Evaluación 39
4.3. Discusión 41

5. Caso de estudio 43
5.1. Detección de ruta anómala 43
5.1.1. Base de Datos 43
5.1.2. Extracción de características y agrupamiento 44
5.1.3. Descubrimientos 45

6. Eficiencia en Tiempo y Memoria 47

7. Conclusiones y Trabajos Futuros **51**

7.1. Conclusiones 51

7.2. Trabajos futuros 52

Bibliografía **57**

Lista de Tablas Páginas

4.1. Resultados cuantitativos. 40

Lista de Figuras Páginas

2.1. Histograma uniforme 18

2.2. Histograma simétrico 18

2.3. Histograma bimodal 19

2.4. Histograma de probabilidad 19

2.5. Representación gráfica de Multilevel Discrete Wavelet Decomposition (MDWD). 22

2.6. Conjunto de trayectorias 27

3.1. Descripción general del modelo propuesto. 32

3.2. Normalización de trayectorias 34

3.3. Descomposición de Trayectoria. 35

4.1. Comparación de la performance con MDWD y DFT utilizando el conjunto de datos CROSS 40

5.1. Detecciones de trayectorias anómalas en el conjunto de datos *Traffic Flow*. 45

6.1. Gráficos de los valores promedio de memoria y de tiempo de ejecución. 48

Lista de Algoritmos Páginas

1. Pseudocódigo de Affinity Propagation. 23

Introducción

Los dispositivos electrónicos y las conquistas del software están transformando el mundo que nos rodea, brindan asistencia a las personas y permiten el crecimiento de nuestra economía. Sin embargo, esta transformación digital solo puede brindarnos su máximo potencial, si explotamos el poder de los datos. Actualmente estamos atravesando una época de revolución de los datos, este fenómeno es impulsado no sólo por la abundancia de datos que existe actualmente, sino por las tecnologías que cambian la forma en que las reunimos, en que las almacenamos, las analizamos y transformamos. Una gran parte de estos datos son generados por sensores que se encuentran en dispositivos, máquinas inteligentes, vehículos modernos entre otros. Si bien mantener esta cantidad de datos fue alguna vez costoso y difícil, las capacidades de almacenamiento crecieron y los costos cayeron, es así que los datos almacenados son ahora un recurso renovable. Con la nueva capacidad de reutilizar los datos y darles nuevos propósitos, podemos continuar con su análisis y transformarlos en nuevas formas de producir conocimiento que a su vez nos permitan ahorrar tiempo, dinero e incluso salvar vidas.

La inteligencia artificial ya esta revolucionando nuestras vidas de una manera sorprendente, donde el software esta ayudando a las personas a descubrir respuestas escondidas dentro de una cantidad enorme y creciente de datos. En el futuro se calcula que los dispositivos que están conectados alrededor del planeta nos ayudarán a comprender su comportamiento de mejor manera y así mejorar nuestro entorno. El desafío se centra en la extracción de conocimiento de los datos para ponerlos en funcionamiento, valiéndonos de nuestro ingenio para entender los valiosos aprendizajes que guardan. Esta capacidad de procesar los datos para transformarlos en conocimientos, y los conocimientos en respuestas, es la que nos permite obtener soluciones para una mejor toma de desiciones, así como también otros considerables desafíos de la actualidad. Se sabe hoy en día, según un cálculo aproximativo, los científicos de datos pueden pasar entre el cincuenta y el ochenta por ciento de su tiempo preparando datos digitales rebeldes antes de que puedan ser explorados para encontrar piezas o patrones útiles.

A pesar del avance acelerado del proceso de digitalización de la información y de la automatización de los procesos, se siguen descubriendo nuevas y mejores formas de extracción de conocimientos o patrones dentro de estos oseamos de datos, haciendo que la extracción automática del conocimiento siga siendo hoy en día desafiante y necesario.

El presente trabajo presenta una metodología para detectar trayectorias anómalas en función de sus características morfológicas. Para ello, seguimos dos etapas: (1) análi-

sis comparativo del desempeño de dos descriptores para agrupar trayectorias similares, y (2) detección de anomalías de trayectoria en función de sus similitudes. Definimos las transformadas Discrete Fourier Transform (DFT) y Multilevel Discrete Wavelet Decomposition (MDWD) como descriptores de trayectoria para generar características y posteriormente detectar anomalías. Nuestros experimentos enfatizan la medida del desempeño en la descripción dentro del espacio de características de coeficientes usando aprendizaje no supervisado, específicamente técnicas de agrupamiento, para crear subconjuntos e identificar aquellos elementos irregulares.

En el trabajo de tesis se define performance como el rendimiento que tiene el descriptor utilizando las métricas de tiempo, memoria, presición y la curva Receiver Operating Characteristic (ROC). Las implicaciones del estudio demuestran que es posible utilizar descriptores en trayectorias para la detección automática de anomalías y el uso de aprendizaje no supervisado para segmentar la información requerida. En cuanto a la hipótesis de esta tesis, se sostiene que el descriptor Wavelet es mejor que el descriptor Fourier, en la detección de trayectorias anómalas. Finalmente, el rendimiento de nuestro estudio y el análisis comparativo se han demostrado a través de múltiples experimentos. Presentamos algunos resultados cuantitativos usando conjuntos de datos sintéticos así como análisis cualitativos a través de un estudios de caso, considerando conjuntos de datos reales que dejan evidencia de nuestra contribución.

El trabajo de investigación esta organizado de la siguiente manera: El Capítulo 1 presenta una introducción al trabajo de investigación así como los aspectos generales que describen de manera general a la tesis. Esto incluye los objetivos generales, objetivos específicos y contribuciones. Por otra parte en la sección 1.2 se presentan trabajos relacionados, estos definidos en la tesis como antecedentes, mostrando que estos trabajos están relacionados con nuestro problema a estudiar. En el Capítulo 2 se presenta un marco teórico sobre temas que son de interés y servirán para afianzar nuestros conocimientos. De esta forma se establece una base teórica para el desarrollo de la propuesta. En el Capítulo 3 se presenta la propuesta del proyecto de tesis, se comienza con una visión general para luego detallar cada uno de los procedimientos propuestos. Para finalizar este capítulo, se presentan recursos y complementos interactivos propuestos que ayudan al entendimiento del la propuesta. En el Capítulo 4 se presentan pruebas y resultados producto de nuestros experimentos. Esto con el objetivo de evidenciar la eficiencia y la utilidad de nuestro abordaje. En el Capítulo 5 se presenta un caso de estudio con una base de datos real, para demostrar la usabilidad de nuestro abordaje. En el Capítulo 6 se presenta un análisis de eficiencia en tiempo de ejecución y memoria consumida por los descriptores a fin de mostrar: el valor, la utilidad y los modestos recursos que nuestro abordaje requiere. Finalmente en el Capítulo 7 se presenta conclusiones relacionadas a nuestros objetivos específicos así como trabajos futuros relacionados.

Capítulo 1

Aspectos Generales

1.1. El problema de Investigación

Entender el comportamiento de trayectorias puede llegar a ser un problema retador debido a la incomprensión de sus características en espacio-tiempo. Las trayectorias cumplen un papel importante en diferentes campos como: en el flujo de trafico (Kim & Mahmassani, 2015a), en la meteorología para el estudio de las corrientes de aire (Powell & Aberson, 2001), para el pronóstico del tiempo (Pang & Liu, 2020), para el planeamiento de vuelos (Paul *et al.*, 2017), para el seguimiento de tornados, también en vídeos de seguridad (Quispe Torres *et al.*, 2019), tráfico marítimo, flujo de actividades (Morris & Trivedi, 2011), seguimiento de animales (Wisdom *et al.*, 2004), deportes (Turchini *et al.*, 2015) entre muchos otros.

En el campo de las anomalías, el comportamiento anómalo puede indicar importantes objetos y eventos en una amplia variedad de dominios (Laxhammar & Falkman, 2014). Sin embargo, este análisis no es un problema trivial debido a la secuencia de análisis, complejidad de la morfología, y la calibración de los parámetros de los algoritmos.

La detección de trayectorias anómalas es un problema importante porque permite identificar trayectorias que pueden indicar actividades ilegales y adversas. Por ejemplo; en vídeo vigilancia, podría indicar agresión personal, robo y sabotaje de infraestructura. Sin embargo, ésta no es una tarea trivial; ya que los algoritmos tienen que enfrentarse a diferentes problemas, entre ellos, por ejemplo esta la limpieza del ruido de las trayectorias y la extracción de información semántica. La información semántica implica experimentación y estudios para transformar movimientos sin significado a otro tipo de representaciones (Parent *et al.*, 2013) con mayor significado. Además, la falta de métricas exactas para medir la calidad de un extractor semántico dificulta el estudio de estas trayectorias.

En cuanto al campo de la minería de datos, las trayectorias pueden estar presentes en diferentes formatos, por ejemplo como series-temporales. Actualmente las

series-temporales presentan problemas retadores para su análisis. Por ejemplo las series-temporales son muy utilizadas en la bolsa de valores para mostrar comportamientos y hacer predicciones.

Por otro lado, hoy en día el uso de smartphones es cada vez más frecuente, haciendo posible el rastreo de personas mediante dispositivos Global Positioning System (GPS) incrustado en los mismos. Cabe recalcar que las trayectorias generadas por GPS son una secuencia consecutiva de coordenadas georreferenciadas (latitud, longitud) con sus respectivos períodos de tiempo en los que se tomó cada muestra. El incremento de estos dispositivos así como su bajo precio a hecho que la tendencia a adquirir estos dispositivos experimente un crecimiento. Debido al crecimiento de estos dispositivos georreferenciados, se hace cada vez más interesante el análisis de estos datos para su exploración y su posterior generación de conocimiento. Es también posible identificar comportamientos anómalos de vehículos con las trayectorias generadas por estos dispositivos.

El termino *descriptor* es mayormente utilizado en el área de procesamiento de imágenes, del cual se hacer referencia para el estudio. El presente trabajo de investigación compara dos descriptores para la detección de trayectorias anómalas tomando como característica principal la morfología. Además, las metodologías propuestas presentan experimentos para verificar que los descriptores mejoran el proceso de análisis de trayectorias. Estos experimentos se validarán mediante comparaciones de rendimiento teniendo en cuenta algunos conjuntos de datos de la literatura. Específicamente, el objetivo es analizar el rendimiento de la detección automática de anomalías en trayectorias utilizando el aprendizaje no supervisado, tomando como descriptores principales la Transformada Discreta de Fourier y la Descomposición Multinivel Discreta de Wavelet. Además, aplicamos nuestra metodología sobre un conjunto de datos de rutas de buses de transporte público urbano, analizando la detección de anomalías en un escenario realista. Finalmente, también hemos detallado la arquitectura de software y hardware que utilizamos para todo nuestro proceso analítico, mostrando la utilidad del método.

1.1.1. Descripción del Problema

Como se mencionó el estudio se centra en el análisis comparativo de dos descriptores para trayectorias, esto en el campo de la detección de anomalías. Además de comprobar si estos descriptores mejoran el procesamiento de trayectorias de manera general. Estas propuestas serán validadas haciendo comparaciones con bases de datos encontradas en la literatura.

Dentro del estudio, una vez aplicado los descriptores, posterior a este proceso se aplicarán métodos de aprendizaje no supervisados. Los métodos de aprendizaje no supervisados nos permitirán segmentar la información de anomalías de manera automática. También serán presentados análisis comparativos de la eficacia de los enfoques presentados a través de experimentos y observaciones en distintos conjuntos de datos que dejan evidencia del aporte del estudio.

El problema de partida se centra en determinar cual es la performance o rendimiento de las transformadas de Wavelet y Fourier como descriptores de trayectorias, así mismo determinar cual de estos descriptores presenta mejor rendimiento, teniendo en cuenta que dentro de la literatura la utilización de estos métodos para propósitos de descripción u obtención de características no esta ampliamente estudiado y existen pocos trabajos relacionados para este fin.

Las grandes cantidades de datos, el ruido que tienen estas, la complejidad con las que hay que lidiar; es decir, con dimensiones como la espacial y la temporal, son algunas de las dificultades encontradas en este estudio. Aportes como el descubrimiento sistemático de agrupaciones espaciales de flujos de tráfico en toda una red son algunos aportes que pueden proporcionar este tipo de estudio; y también en el agrupamiento de trayectorias sin información adicional; es decir, que no requiere información adicional del entorno o algún demarcado previo al procesamiento.

Por último, se sabe que la información semántica que éstos datos contienen, llega a ser difícil de extraer, ya que el significado de cada trayectoria puede variar debido al contexto. Otro aspecto a considerar, es el hecho de la falta de métricas exactas para medir la calidad de un extractor semántico caracteristicas o información en trayectorias. Estas son algunas de las respuestas que motivan la investigación.

1.1.2. Identificación del Problema

¿Cúal es el descriptor, Wavelet o Fourier, que obtenga la mayor performance en la detección de trayectorias anómalas?

1.2. Antecedentes

La literatura concerniente al análisis de trayectorias es extensa. Para contextualizar mejor nuestro abordaje, dividimos esta sección en cuatro partes, consideramos los conjuntos de datos utilizados en nuestra experimentación, así como, una introducción al análisis de trayectorias, el procesamiento de las mismas y la detección de anomalías.

1.2.1. Introducción al análisis de trayectorias

Kong *et al.* (2018) clasifica los datos de trayectorias como implícitas y explicitas basadas en la continuidad y estructura que estas presentan.Aquellos datos de trayectoria explicitas proveen información de tiempo y locación, ademas de una buena estructura y una continuidad espacio temporal solida. Las trayectorias generadas por dispositivos GPS son las mas representativas en esta categoría. Por otro lado, los

datos de trayectoria implícitas tiene una carencia de continuidad espacio temporal, sub-categorizando esta clase en aquellas basadas en señales, basadas en sensores y basados en datos de redes.

Este estudio mencionado e introduce un resumen de aplicaciones y servicios que usan trayectorias, presentando así una clasificación de trayectorias basadas en aplicaciones, también este estudio presenta servicios de sistemas de recomendación que utilizan trayectorias en sus estudios. Para contextualizar, de acuerdo a este estudio, parte de los conjuntos de datos utilizados en nuestro estudio pertenece a la sub-categoría de datos basados por sensores, debido a que estos datos son generados por el monitoreo de personas por cámaras de seguridad.

1.2.2. Procesamiento de trayectorias

Dependiendo de la entidad que origine las trayectorias, estas estarán sometidas a un conjunto finito de clases o tipos de trayectorias, desde movimientos regulares hasta altamente variables. El modelamiento de trayectorias es el primer y mas desafiante paso, en el tratamiento de trayectorias.

Dentro de la literatura, existen diferentes formas de tratar las trayectorias; por ejemplo, el algoritmo denominado TRACLUS (Lee *et al.*, 2007), este algoritmo procesa las trayectorias usando los segmentos que lo conforman, creando con esta información un resumen de todos estos. Por otro lado, algunos estudios restringen las trayectorias a una red de transporte o senderos; esto referido al movimiento de vehículos seguido a la red de transporte.

En esta categoría, NETSCAN (Lee *et al.*, 2007), NNCluster (Roh & Hwang, 2010), y NEAT (Han *et al.*, 2012) muestran estudios restringidos o sometidos a una red de caminos; es decir que estan sometidos o restringidos a dicha red, NETSCAN y TRACLUS procesan trayectorias desde un enfoque de segmentación de trayectorias sin considerar características o patrones que se repiten en diferentes partes de la trayectoria (características de mayor nivel).

En cambio los trabajos de Naftel & Khalid (2006), de Khalid & Naftel (2010), Annoni & Forster (2012) modelan trayectorias capturando información presente en toda la trayectoria, es decir sin crear segmentos de recta.

Por otro lado el proceso de agrupamiento (aprendizaje no supervisado) no es nuevo dentro de la literatura, este también fue utilizado con el modelamiento de trayectorias de vehículos como grafos (Rinzivillo *et al.*, 2008), y también fue utilizado para generar visualizaciones de trayectorias (Guo *et al.*, 2010).

1.2.3. Detección de anomalías

En este trabajo, para detectar anomalías dentro de un espacio de vectores de características, utilizamos el abordaje basado en distancias (Distance-Based Methods) segun Zhang *et al.* (2020). Aquellas trayectorias que se encuentran a una larga distancia desde un conjunto de trayectorias son recuperadas como anómalas, utilizando el agrupamiento para crear los grupos que están conformados por trayectorias que son similares.

1.2.4. Conjunto de datos utilizados

Piciarelli *et al.* (2008) creo un algoritmo para generar conjuntos de datos de trayectorias sintéticas. Este algoritmo puede generar mil subconjuntos de trayectorias que son automáticamente generadas, cada trayectoria esta conformada por diez y seis puntos. Estos conjuntos de datos se utilizan en diversos estudios academicos, debido a que este estudio es uno de los primeros trabajos que aborda la detección de trayectorias anomalas, y ademas que comparte el conjunto de datos que generó. Años después, Laxhammar & Falkman (2014) presentan nuevos resultados considerando el algoritmo y los conjuntos de datos generados por Piciarelli. Este estudio enfatiza el análisis secuencias incompletas de trayectorias, que el autor lo denomina real-time learning" ó aprendizaje en tiempo real, basado en actualizaciones incrementales del conjunto de entrenamiento.

Este estudio propone e implementa un detector de anomalias llamado Sequential Hausdorff Nearest-Neighbor Conformal Anomaly Detector (SHNN-CAD) para un aprendizaje en tiempo real y una detección secuencial de detección de anomalías en trayectorias, obteniendo una performance competitiva para la clasificación con un minimo de configuración de hiperparametro.

Ergezer & Leblebicioğlu (2016) presentan un descriptor de trayectorias basada en una matriz de covarianza para detectar trayectorias anómalas usando Nearest-Neighbors (NN) y Space-Representation (SR), además del uso del agrupamiento espectral para la percepción de actividad.

Este estudio también utiliza los datos sintéticos creados por Piciarelli *et al.* (2008) como parte de sus resultados y también ejecuta experimentos con datos reales de la universidad de San Diego California (UCSD), y MIT Parking Lot (Wang *et al.*, 2011) para la detección de anomalías en vídeos.

De la misma manera, Sillito & Fisher (2008) proponen una nueva estructura para detectar trayectorias anormales considerando el comportamiento de transeúntes en términos de movimiento de trayectoria. Esta estructura construye un clasificador One-Class, que es basado en probabilidades usando la distribución de Gauss (Gaussian distribution). Además, ellos condujeron experimentos usando conjuntos de datos etiquetados y no etiquetados, usando dos conjuntos de datos como CAVIAR INRIA (Labs,

2004) y Capark (Dee & Hogg, 2004).

1.3. Objetivos

1.3.1. Objectivo General

Comparar la performance de los descriptores Wavelet y Fourier, para la detección de anomalías en trayectorias.

1.3.2. Objetivos Específicos

1. Desarrollar una metodología que identifica anomalías basadas en la morfología de trayectorias utilizando Discrete Fourier Transform (DFT) y Multilevel Discrete Wavelet Decomposition (MDWD). Se utilizará esta metodología para obtener resultados y hacer comparaciones entres estos dos descriptores.
2. Verificar el rendimiento de métodos de aprendizaje no supervisados como Affinity Propagation (AP) en la detección de trayectorias anómalas.
3. Demostrar la usabilidad de la metodología propuesta, en la detección de trayectorias anómalas utilizando bases de datos presentes en otros estudios académicos.

1.4. Hipótesis

La pregunta de partida planteada sugiere la hipótesis de investigación que a continuación se relaciona:

H_i : "La performance del descriptor de Wavelet será mejor al de Fourier en la detección de trayectorias anómalas".

La hipótesis planteada busca pronosticar o predecir un dato o valor en dos variables que se van a medir u observar, esto debido al alcance *descriptivo* de nuestra investigación según Hernández-Sampieri & Torres (2018).

Se define la hipótesis alternativa que se relaciona con la pregunta de partida:

H_a : "La performance del descriptor de Wavelet será igual al de Fourier en la detección de trayectorias anómalas".

Se define la hipótesis nula que se relaciona con la pregunta de partida:

H_0 : "La performance del descriptor de Wavelet no es mejor al de Fourier en la detección de trayectorias anómalas".

1.4.1. Definición de Variables

Según Hernández-Sampieri & Torres (2018) una variable es una propiedad que puede fluctuar y cuya variación es susceptible de medirse u observarse. Para el presente trabajo de investigación se identifican las siguientes variables:

X_1 : Descriptor de trayectorias utilizando DFT.

X_2 : Descriptor de trayectorias utilizando MDWD.

Ambas variables son independientes una de la otra y se asegura de que las variables pueden ser medidas, observadas, evaluadas o inferidas, es decir, que de ellas se pueden obtener datos en la realidad; además, ambas variables son utilizadas para el planteamiento de la hipótesis.

1.5. Justificación

Justificación Teórica. Dentro de la literatura se encuentran pocos estudios en artículos científicos que exploren a profundidad el concepto de descriptor de trayectorias que sean sensible a alteraciones en la morfología de las trayectorias. Como se muestra en la Sección 1.2, los estudios relacionados usan varios métodos para extraer características diferentes desde trayectorias. Sin embargo; el uso de las transformadas de Wavelet y de Fourier, no presenta estudios profundos para este fin.

Además según Xu *et al.* (2015), quien realizó experimentos en el área de agrupamiento de trayectorias, afirma que apesar de que los puntos de las trayectorias nos dan información de posición, estos puntos por si solos no proveen suficiente información, justificando la necesidad de aplicar procedimientos previos al agrupamiento de trayectorias, para así evitar perdidas de información. Se cita textualmente a *Xu*, quien dice lo siguiente: "A pesar de que las trayectorias por si mismas recolectan información de posición de un determinado objeto a través del tiempo y contienen información significante para un proceso de agrupamiento, computar directamente el agrupamiento en estas posiciones provee pobres resultados". Nuestro estudio muestra resultados con el agrupamiento de trayectorias realizando un pre-procesamiento, demostrando así, que un descriptor de trayectorias puede ser un diferenciador considerable para mejorar resultados.

Cabe mencionar que la utilización de descriptores, tiene una analogía con la generación de espacios latentes. Para este caso se estaría hablando de un espacio donde la morfología de una trayectoria es una característica presente en el definido espacio. También en la literatura se habla de espacios de similaridad. Los resultados de este trabajo, contribuirían en alguna medida a estos conceptos.

Justificación práctica. La demanda de procesamiento y extracción de conocimiento a partir de datos ha venido en aumento en los últimos años, lo que ha generado la necesidad de automatizar estos procesos para obtener resultados precisos que permitan tomar mejores decisiones. Lamentablemente, en la actualidad, los datos no están siendo aprovechados eficientemente, lo que resulta en la pérdida de valioso conocimiento que podría prevenir errores y accidentes, así como mejorar la toma de decisiones. Por lo tanto, se requiere una investigación continua sobre métodos que faciliten el análisis de grandes cantidades de datos y permitan su utilización de manera efectiva. Los resultados de la investigación sirven como aporte para demostrar que es posible utilizar *coefficient feature spaces* para detectar trayectorias anómalas, tomando un menor tiempo de ejecución en comparación a un proceso de entrenamiento de una red neuronal y mostrando resultados competitivos con trabajos dentro de la literatura.

Las anomalías pueden darnos información de un evento extraño en el mundo real. Detectarlos de manera automática nos permite tener un mejor control de determinadas situaciones en las cuales ocurren estas anomalías. Este trabajo recoge información semántica que se incluye en este proceso de detección de anomalías. Se sabe, dentro de la literatura, que la extracción de información semántica es un problema desafiante y latente. Para este estudio llamaremos información semántica a aquella información extraída de los conjuntos de trayectorias, la información semántica permite diferenciar trayectorias anormales de otras normales.

Justificación metodológica. Se sabe que nuestra investigación tiene un enfoque cuantitativo, como se explicará en la Sección 1.8, esto debido a los siguientes factores:

- Objetividad: La metodología cuantitativa se basa en datos numéricos y estadísticas, lo que permite que los resultados obtenidos sean más objetivos y menos subjetivos que en una investigación cualitativa.
- Validación empírica: La metodología cuantitativa se enfoca en la recolección de datos empíricos, lo que permite validar las hipótesis de la investigación de manera más rigurosa.
- Generalización de resultados: La metodología cuantitativa se enfoca en la recolección de datos a gran escala, lo que permite generalizar los resultados a una población más amplia.
- Precisión: La metodología cuantitativa utiliza instrumentos de medición estandarizados y objetivos, lo que permite obtener resultados más precisos.
- Eficiencia: La metodología cuantitativa permite recolectar y analizar grandes cantidades de datos de manera más eficiente que en una investigación cualitativa.

Cada una de estas razones concuerdan con las caracteristicas presentes en esta tesis de investigación, concordando con los objetivos de la investigación y la naturaleza del fenomeno estudiado.

1.6. Métodos de Evaluación

El proceso de evaluación de nuestro abordaje se divide en dos partes. Esto debido a la naturaleza de los conjuntos de datos utilizados. Se realiza un proceso de evaluación para los conjuntos de datos que fueron generados sintéticamente, y también se realiza un proceso de evaluación para conjuntos de datos reales. Se describe brevemente cada uno de estos métodos así como el porque de la utilización de los mismos en cada caso.

La primera parte de nuestros experimentos se realizó con conjuntos de datos sintéticos. Para esta evaluación utilizaremos dos métricas, como son precisión promedio y Area Under the Curve (AUC), esta última generada por la curva ROC. Debido a que se utilizan las bases de datos generados por Piciarelli *et al.* (2008), Laxhammar & Falkman (2014); y a fin de realizar una evaluación exhaustiva, se utilizará la métrica denominada precisión promedio. Estas dos primeras bases de datos nos permiten extraer múltiples valores de precisión para diferentes conjuntos de trayectorias, luego de extraerlas, estos valores son promediados a fin de tener una idea general. Estos resultados son presentados en la Tabla 4.1.

En cambio para la tercera base datos denominada CROSS, que esta generada por un solo conjunto de trayectorias, se utiliza la métrica denominada AUC. Debido a los diferentes umbrales que se tiene para definir anomalías, se decide evaluar el rendimiento de esta base de datos utilizando la curva ROC. Cada punto perteneciente a la curva ROC es obtenida con un valor diferente de umbral, estos umbrales puede definir diferentes valores para True Positive Rate (TPR) y False Positive Rate (FPR). Otra justificación para la utilización de estas métricas son los trabajos relacionados, ya que por ejemplo son utilizados en Mora *et al.* (2020) y Dias *et al.* (2020). La curva ROC permite: visualizar la precisión del detector, facilitar la comparación de uno a más abordajes y también permitir reconocer la importancia de las decisiones del umbral elegido (Moumena, 2016). Estos conceptos son profundizados en el Capitulo 2.

La segunda parte de los experimentos fueron realizados con conjuntos de datos reales. Para ello se evalúa el rendimiento en tiempo de ejecución y uso de memoria de los dos descriptores propuestos. Se muestran medidas realizadas a cada experimento, siendo posible hacer comparaciones de eficiencia entre los descriptores propuestos. Luego de la presentación de estas mediciones se realiza un análisis de complejidad algorítmica.

1.7. Alcances y Limitaciones

1.7.1. Alcances

Los alcances de este estudio se pueden resumir en:

- La detección de trayectorias anómalas utilizará la morfología de cada una de las

trayectorias como característica a considerar.

- Las trayectorias de diferentes longitudes y de diferente número de puntos son tratadas por igual sin necesidad de ninguna eurística.
- Los puntos que conforman las trayectorias recolectadas pueden ser tomados en diferentes intervalos de tiempo.
- Se define un problema el cual es abordado y validado con bases de datos reales y sintéticos, que en su mayoría se encuentran en otros estudios académicos.
- Se presenta un caso de estudio en el cual se analizan trayectorias generadas por un Sistema de Transporte Inteligente (STI) que esta implementado en la ciudad de Cusco (Alvarez Mamani, 2018), los datos generados por dicho estudio son utilizados en nuestro proceso de evaluación.

1.7.2. Limitaciones

Algunas limitaciones de este estudio pueden resumirse en:

- La detección de trayectorias anómalas depende de la frecuencia con la que aparecen en el conjunto de datos definido, esto debido al método de agrupamiento automático que se utiliza.
- Es necesario que las trayectorias a procesar no contengan ruido, ya que esto altera la morfología real de una trayectoria.

1.8. Metodología de Investigación

Tanto el enfoque como el diseño de investigación que tendrá nuestra investigación son extraídas de acuerdo al trabajo de Hernández-Sampieri & Torres (2018), que es en la que nos basamos para describir los *alcances* del trabajo.

Desde un punto de vista holístico, el presente trabajo tiene un enfoque ***cuantitativo***, debido a que nuestra investigación es secuencial y probatoria además de recolectar datos para probar hipótesis con base en la medición numérica y el análisis estadístico, con el fin de establecer pautas y probar teorías. Según este autor dentro de la investigación existen *alcances*, y estos no deben considerarse como "tipos", ya que mas que ser una clasificación, constituyen un continuo de "casualidad" que puede tener un estudio.

De acuerdo a Hernández-Sampieri & Torres (2018), una misma investigación puede incluir diferentes alcances, asimismo, es posible que una investigación se inicie como *exploratoria* y después llegue a ser *correlacional* y aun *explicativa*. Teniendo en cuenta esto se confirma que nuestra investigación es ***exploratoria y descriptiva***, debido a

que la meta de la investigación *descriptiva* consiste en describir fenómenos, situaciones, contextos y sucesos; esto es, detallar cómo son y se manifiestan. Con los estudios descriptivos se busca especificar las propiedades, las características y los perfiles de objetos o cualquier otro fenómeno que se someta a un análisis. Es decir, únicamente pretenden medir o recoger información de manera independiente o conjunta sobre los conceptos o las variables a las que se refieren, su objetivo no es indicar cómo se relacionan éstas. Los estudios *descriptivos* son útiles para mostrar con precisión los ángulos o dimensiones de un fenómeno, suceso, comunidad, contexto o situación (Hernández-Sampieri & Torres, 2018).

La metodología que se presenta está en concordancia a los pasos que se siguieron en el proceso de investigación. La secuencia de fases que se sigue es la siguiente:

1. **Revisión de la literatura:** Se realiza una búsqueda dentro de la literatura en busca de aportes que se hicieron para el procesamiento de trayectorias. Esto incluye estudiar e implementar métodos para describir trayectorias que se encuentran en el estado de arte. También, encontrar bases de datos que permitan evaluar el enfoque propuesto.

2. **Obtención de datos para el estudio:** Consiste en la obtención de datos relacionados al estudio, esto incluye el procesamiento y limpiado de los datos encontrados en el proceso de revisión de la literatura. Además de generar un conjunto de datos especifico para nuestro caso de estudio.

3. **Investigación de la solución:** Buscar la mejor combinación entre descriptores y métodos de aprendizaje no supervisado (clustering methods). Se realizar una evaluación cualitativa para la evaluación de las combinaciones encontradas.

4. **Diseño e implementación de la solución:** Consiste en la implementación del detector de anomalías con los métodos seleccionados. La solución consta de dos módulos principales que son:

    a. El extractor de características: Consiste de métodos de extracción de características para trayectorias. Las características generadas por estos métodos serán analizadas con la ayuda de herramientas de visualización de datos.

    b. El Agrupador de trayectorias: Consta del funcionamiento correcto de los agrupadores de características, considerando la cantidad de dimensiones que estas pueden llegar a tener. También consiste en la búsqueda de los *hiperparametros* para la solución optima de generación de grupos.

5. **Evaluación y análisis de los resultados:** Consiste en la evaluación del método propuesto. Esto conlleva a la validación de la propuesta con todas las bases de datos. Esto incluye una evaluación tanto con datos artificiales como con datos reales. Además de la implementación de los métodos de evaluación seleccionados.

1.9. Contribuciones de esta tesis

1.9.1. Contribuciones en el Ámbito Académico

Durante el periodo del desarrollo del proyecto de investigación, se ha realizado los siguiente aportes científicos:

- Publicación en la conferencia del Centro Latinoamericano de Estudios de Informáctica (CLEI), aprobado y defendido en el mes de Octubre del 2021 con el título "*Trajectory Anomaly Detection based on Similarity Analysis*", Quispe-Torres, GF., Enciso-Rodas, L., Vera-Olivera, H., & Garcia-Zanabria, G., 2021 XLVII Latin American Computing Conference (CLEI), 1-10
  doi: https://doi.org/10.1109/CLEI53233.2021.9639966

- Publicación en la revista científica CLEI electronic journal (CLEIej), se presentó una extensión del trabajo a invitación, el trabajo quedo publicado como mejor paper CLEI 2021 en su categoría, con el título "*A Feature-based Trajectory Anomaly Detection*", GF Quispe-Torres, L Enciso-Rodas, H Vera-Olivera, G Garcia-Zanabria, Special Issue of CLEI 2021 best papers 25 (2), 1-20
  doi: https://doi.org/10.19153/cleiej.25.2.3

1.9.2. Contribuciones Generales

Las principales contribuciones de este trabajo pueden resumirse en:

- Se presenta un análisis comparativos de la performance de los descriptores Wavelet y Fourier, aplicado a la detección de anomalías en trayectorias.
- El desarrollo de un detector de trayectorias anómalas que puede ser aplicado a la identificación de rutas anómalas. Este estudio puede ser aplicado para el control de rutas en sistemas de transporte público y así proporcionar una alarma de control de rutas entre otras aplicaciones.
- Se presenta el tema de la morfología en trayectorias como característica, este tema presenta poca atencíon dentro de la literatura, agregando así un aporte en este tema en específico.
- La utilización de descriptores conjuntamente con un método de aprendizaje no supervisado que permiten procesar trayectorias y detectar anomalías automáticamente. Esta combinación adecuada puede ser considerada como un aporte.

Capítulo 2

Marco Teórico

En esta sección se describe conceptos relacionados al procesamiento de trayectorias y detección de trayectorias anómalas.

2.1. Datos de Trayectoria

Las siguientes representaciones de trayectorias fueron inspiradas por el estudio de Panagiotakis *et al.* (2011).

Definition 2.1.1. *Un punto p es una tupla (x, y, t), donde x e y son las posiciones por donde se mueve el objeto (latitud, longitud) y t es el lapso de tiempo en el cual fue colectado el punto, donde $k \in \mathbb{N}$.*

$$p_k = (x_k, y_k, t_k). \tag{2.1}$$

Una lista de puntos ordenadas en el tiempo forman una trayectoria T_i.

Definition 2.1.2. *Una trayectoria T_i es una tupla (tid_i , $\{p_1, p_2, ..., p_K\}$), donde tid_i es el identificador y $t_1 < t_2 < t_3 < ... < t_K$ en una secuencia de puntos $\{p_1, p_2, ..., p_K\}$, donde $\{i, K\} \in \mathbb{N}$.*

$$T_i = (tid_i, \{p_k\}_{k=1:K}). \tag{2.2}$$

En general, en una trayectoria queda representada el tiempo y la ubicación de un objeto en movimiento que puede ser rastreado. Una trayectoria simple puede incluir muchos puntos y pueden tener cualquier longitud para un simple objeto en movimiento. Básicamente, la longitud representa la ubicación que el objeto a recorrido desde un punto "A" hacia otro punto "B" antes de que se detenga.

Por otro lado para nuestro estudio definiremos una sub-trayectoria como:

Definition 2.1.3. *Una sub-trayectoria T'_s esta dado por un conjunto de números.*

$$T'_s = \{p_k\}_{k=1:K} \tag{2.3}$$

donde T'_s es un conjunto de puntos $p_k = (c_k, t_k)$, t_k es el instante de tiempo en el cual la componente c_k es colectada y $c_k \in w_x \vee c_k \in w_y$, estos dos últimos espacios definidos en la ecuación 3.1 y 3.2 respectivamente.

2.2. Números Complejos

Un número como $5 + 3i$ se llama número complejo. Es la suma de dos términos donde cada uno de estos términos puede ser cero. El término real que no contiene i, se llama parte real mientras que el coeficiente de i es la parte imaginaria. Por lo tanto, la parte real de $5 + 3i$ es 5 mientras que la parte imaginaria es 3. Un número es real cuando el coeficiente de i es cero, y a su vez es imaginario cuando la parte real es cero, por ejemplo: $5 + 0i = 5$ es real mientras que $0 + 3i = 3i$ es imaginario.

Habiendo introducido un número complejo, es necesario definir las formas en que se pueden combinar, es decir, la suma, la multiplicación o la división. Esto se denomina el álgebra de los números complejos. Verás que, en general, se procede como en números reales, pero usando $i^2 = -1$ donde corresponda. Para ello se explica cuando dos números complejos son iguales: sean $a + bi$ y $c + di$ dos números complejos iguales, entonces sus partes reales e imaginarias son iguales; es decir, si $a + bi = c + di \Rightarrow a = c \wedge b = d$.

2.2.1. Adición y Sustracción de Números complejos

La **suma** de números complejos se define sumando por separado partes reales e imaginarias: sea $z = a + bi$ y sea $w = c + di$ entonces

$$z + w = (a + c) + (b + d)i$$

y esto del mismo modo para la resta.

2.2.2. Multiplicación y División de Números Complejos

Para la **multiplicación** de números complejos se tiene la siguiente fórmula en general: sea $z = a + bi$ y sea $w = c + di$ entonces

$$z \times w = (a + bi)(c + di) \Rightarrow z \times w = ac - bd + (ad + bc)i$$

donde $i^2 = -1$.

Para la **división** de dos números complejos utilizaremos el conjugado de un número complejo así como la propiedad de los Binomios Conjugados; es decir:

- Sea el número complejo $z = a + bi$, su complejo conjugado esta dado de la forma $\bar{z} = a - bi$. El complejo conjugado de un número complejo se obtiene cambiando el signo de la parte imaginaria.

- Por otra parte, el producto de binomios conjugados, es decir la suma de dos cantidades multiplicadas por su diferencia es igual al cuadrado de la primera cantidad menos el cuadrado de la segunda, donde se cumple la fórmula:

$$(a+b)(a-b) = a^2 - b^2$$

La **división** de dos números complejos es similar a la división de dos números reales; es decir, si $z_1 = a + bi$ y $z_2 = c + di$ son números complejos, la división se puede escribir como:

$$\frac{z_1}{z_2} = \frac{a+bi}{c+di}$$

y a su vez este cociente puede ser obtenido utilizando la siguiente fórmula:

$$\frac{z_1}{z_2} = \frac{ac+bd}{c^2+d^2} + i(\frac{bc-ad}{c^2+d^2}) \tag{2.4}$$

esta fórmula puede ser fácilmente demostrara utilizando la conjugada del denominador, el binomio conjugado y que $i^2 = -1$, como se muestra a continuación:

$$\begin{aligned}
\frac{a+bi}{c+di} &= \frac{(a+bi)(c-di)}{(c+di)(c-di)} \\
&= \frac{(a+bi)(c-di)}{c^2-(di)^2} \\
&= \frac{ac-iad+ibc-i^2bd}{c^2-(-1)d^2} \\
&= \frac{(ac+bd)+(bc-ad)i}{c^2+d^2} \\
&= \frac{ac+bd}{c^2+d^2} + \frac{bc-ad}{c^2+d^2}i
\end{aligned}$$

2.3. Histogramas

El histograma es una de las herramientas gráficas más importantes en la práctica estadística. El histograma proporciona una estimación consistente de cualquier función de densidad con muy pocas suposiciones. El histograma en forma pictórica proporciona el resumen gráfico más común de una muestra aleatoria, así como una estimación de la función de densidad de probabilidad subyacente. Los puntos de datos se tabulan en una lista de contenedores separados (Scott, 2015).

2.3.1. Tipos de Histogramas

Los histogramas se pueden clasificar en diferentes tipos según la distribución de frecuencia de los datos. Existen diferentes tipos de distribución, como la distribución

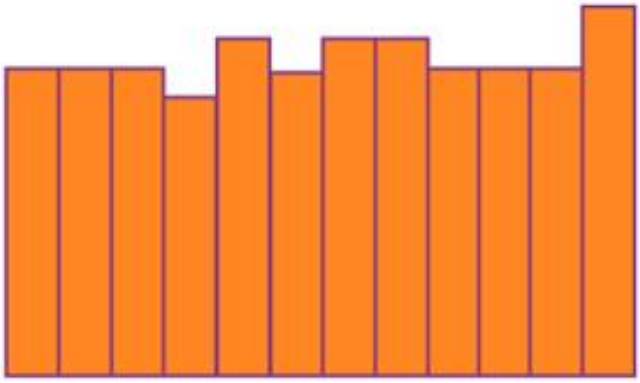


Figura 2.1: Histograma uniforme. Imagen extraída del trabajo de Scott (2015).

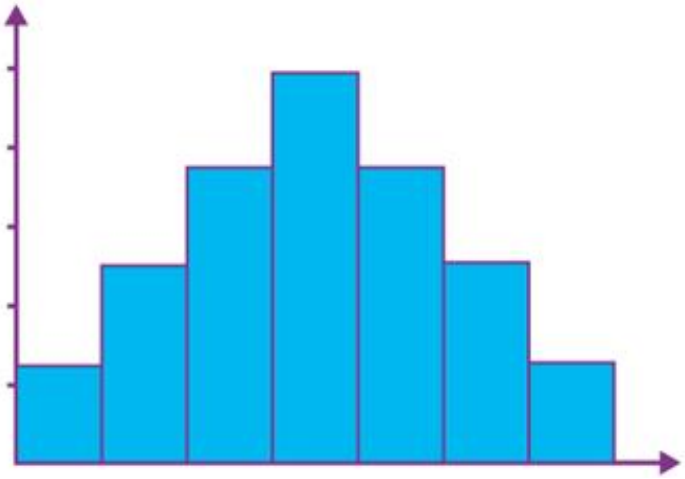


Figura 2.2: Histograma simétrico. Imagen extraída del trabajo de Scott (2015).

normal, la distribución sesgada, la distribución bimodal, la distribución multimodal, la distribución en peine, la distribución en punta marginal, la distribución tangencial, entre otros más (Scott, 2015). Algunos tipos de histogramas son:

Histograma Uniforme. Una distribución uniforme indica que el número de clases presentes es demasiado pequeño y cada clase tiene el mismo número de elementos. Esto puede incluir una distribución con múltiples picos. La Figura 2.1 muestra una distribución uniforme donde no existen picos pronunciados.

Histograma Simétrico. Un histograma simétrico también se llama histograma en forma de campana. Se dice que un histograma es simétrico si es posible dibujar una línea vertical en la mitad del histograma, de tal forma que esta linea pueda definir dos histogramas de igual tamaño y forma con respecto a esta linea. Es decir, tanto el lado derecho del histograma es similar al lado izquierdo, formando una gráfica perfectamente simétrica. Por otra parte a los histogramas asimétricos se le denominan sesgados. La

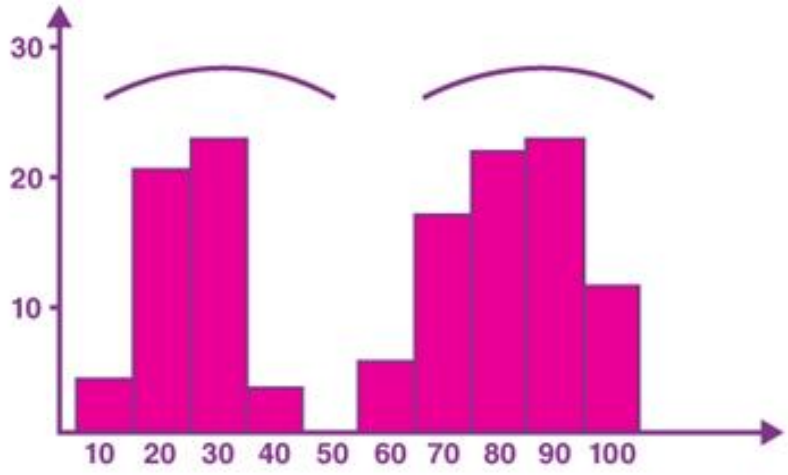


Figura 2.3: Histograma bimodal. Imagen extraída del trabajo de Scott (2015).

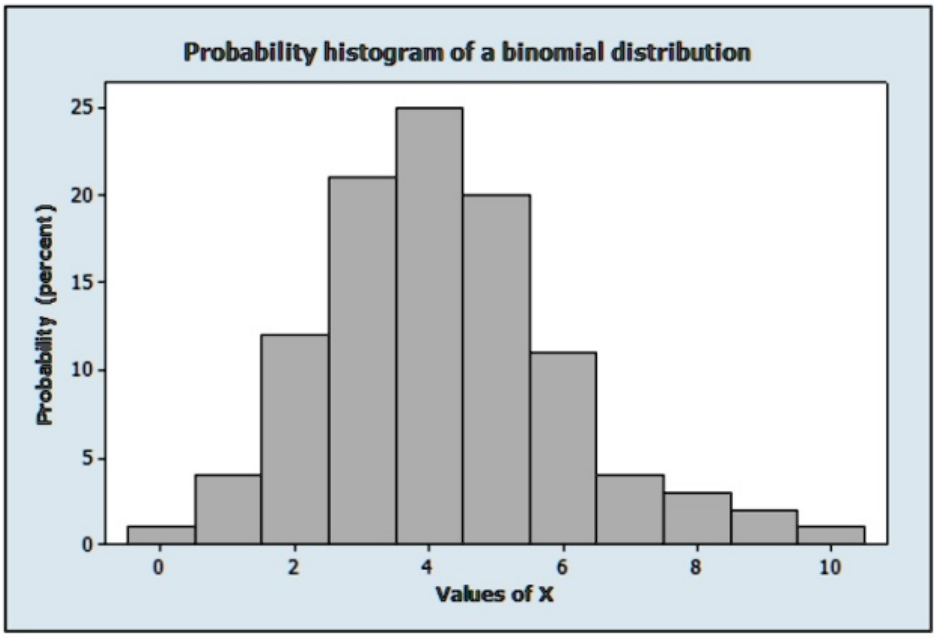


Figura 2.4: Histograma de probabilidad de una distribución binomial.

Figura 2.2 muestra un ejemplo de histograma simétrico.

Histograma Bimodal. Un histograma se llama Bimodal si tiene dos picos; es decir, si los centros de los dos histogramas individuales están lo suficientemente separados de la variabilidad de dos conjuntos de datos. La Bimodalidad resulta en general de una distribución de dos distribuciones normales y sugiere la existencia de datos de dos procesos diferentes. Tiene un valle en el centro del rango de entre sus picos. Es decir, en un histograma bimodal, el valle corresponde a la frecuencia menor entre los dos picos. La Figura 2.3 muestra un ejemplo de un histograma bimodal.

Histograma de Probabilidad. Un histograma de probabilidad muestra una representación gráfica de una distribución de probabilidad discreta. Consiste en un rectángulo centrado en cada valor de x, y el área de cada rectángulo es proporcional a la probabilidad de ese valor. El histograma de probabilidad comienza con la selección de clase. La probabilidad de cada resultado es la altura de las barras del histograma. La Figura 2.4 muestra un ejemplo de un histograma de probabilidad.

2.4. La Transformada de Fourier

La transformada de Fourier de una función $f(t)$ en el dominio del tiempo esta dado por:

$$F(\omega) = \int_{-\infty}^{+\infty} f(t) e^{-j2\pi\omega t} dt, \tag{2.5}$$

donde $j = \sqrt{-1}$ y $F(\omega)$ es la función en el dominio de las frecuencias y la función $f(t)$ puede ser obtenido usando la transformada inversa de Fourier.

2.4.1. La Transformada Discreta de Fourier (DFT)

Como se mencionó la transformada de Fourier es una función matemática que descompone una onda o señal. Esta varia en el tiempo en lo que respecta a la frecuencia y amplitud de las cuales la onda o señal esta compuesta. La salida de la transformada de Fourier esta compuesta por partes de números reales e imaginarios para las frecuencias tanto negativas como positivas. El valor absoluto de las salidas representan las frecuencias de la función original. Debido a esto, las transformadas de Fourier permiten ver a cualquier función como una suma simple de sinusoides.

La transformada discretas de Fourier es un tipo de transformada discreta usada en el análisis de Fourier (Smith *et al.*, 1997; Cooley *et al.*, 1969). La transformada discreta de Fourier pueden ser descrita como muestras de una función en determinadas frecuencias, y esto como entrada requiere de una secuencia discreta finita. Por ejemplo, esta secuencia puede ser generada desde una sección de la señal.

Sea $f(t)$ la señal que es el origen de los datos a describir y sea N el numero de muestras separadas por intervalos de tiempo denotados por $f[0], f[1], f[2], ..., f[N-1]$. La transformada discreta de Fourier de $f(t)$ puede ser definida como:

$$F[k] = \sum_{n=0}^{N-1} f[n] W^{kn}, \tag{2.6}$$

para cada $k = 0, 1, \ldots, N-1$. Donde $W = e^{-j(2\pi/N)}$ y $j = \sqrt{-1}$ el cual es un número imaginario. $F[k]$ son los coeficientes de cada función base en una sumatoria lineal.

2.5. La transformada de Wavelet

La transformada de Wavelet es una función matemática utilizada en el procesamiento digital de señales y compresión de imágenes. En el procesamiento de señales, Wavelets hace posible la recolección de señales débiles separando el ruido exitosamente, lo cual es muy útil, especialmente en el proceso de rayos-X e imágenes por resonancias magnéticas en aplicaciones médicas. Las transformadas de Wavelet y de Fourier representan las señales a través de combinaciones lineales de las funciones base, y ambos descomponen las señales como una superposición de unidades simples desde la cual las señales originales pueden ser reconstruidas. Las transformadas de Wavelet descomponen las señales en Wavelets, y sus funciones base son compactas o finitas en el tiempo. Esta característica permite a las transformadas de Wavelet obtener información de tiempo acerca de la señal además de la información de frecuencia. La transformada de Wavelet tiene un tamaño de ventana que hace variar la escala de la frecuencia. Esta técnica es ventajosa para el análisis de señales que contienen tanto discontinuidades como componentes suaves. Se necesitan funciones de base cortas de alta frecuencia para las discontinuidades, mientras que al mismo tiempo, se necesitan funciones largas de baja frecuencia para los componentes suaves. Los Wavelets son una clase de funciones que se utilizan para localizar una función determinada tanto en el espacio como en la escala.

2.5.1. La Descomposición Multinivel Discreta de Wavelet

Para analizar las señales no estacionarias, es necesario descomponer las señales en unidades localizables, es decir en dominios tanto en tiempo como en frecuencia. Para este propósito, se utiliza MDWD. Según Wang *et al.* (2018), la MDWD es un método de señal discreta basado en Wavelet, la cual puede extraer caracteristicas de frecuencia tiempo multinivel desde una señal, descomponiendo esta como sub-señales de alta y baja frecuencia nivel por nivel.

Para la siguiente explicación utilizamos las letras en negrita como $\mathbf{x}$, $\mathbf{a}$ o $\mathcal{X}$ para denotar vectores y las letras que no están en negrita a, x o l para denotar los escalares. Denotamos la entrada de N ejemplos para la señal como $\mathbf{x} = \{x_0, x_1, \ldots, x_{N-1}\}$, y las sub-señales altas y bajas generadas en el i-th nivel como $x^l(i)$ y $x^h(i)$. En el $(i+1)$-th nivel, MDWD utiliza filtros paso bajo $\mathbf{l} = \{l_1, \ldots, l_k, \ldots, l_K\}$ y un filtro paso alto $\mathbf{h} = \{h_1, \ldots, h_k, \ldots, h_K\}$, $K \ll N$, para convolucionar sub-señales de baja frecuencia del nivel superior como:

$$a_n^l(i+1) = \sum_{k=1}^{K} x_{n+k-1}^l(i) \cdot l_k, \tag{2.7}$$

$$a_n^h(i+1) = \sum_{k=1}^{K} x_{n+k-1}^l(i) \cdot h_k, \tag{2.8}$$

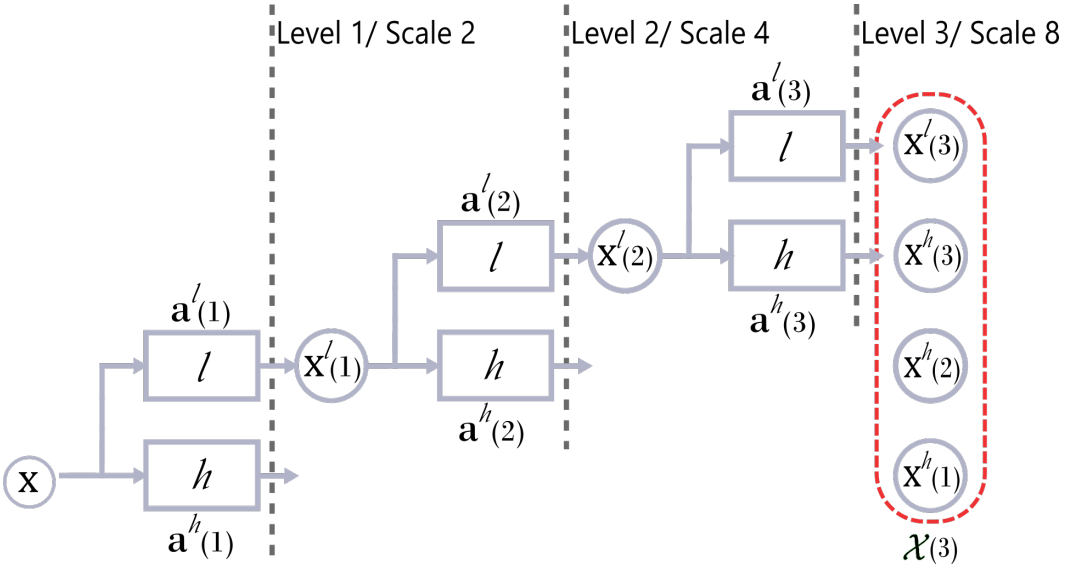


Figura 2.5: Representación ilustrativa de MDWD para $\mathbf{x}$ con tres niveles obteniendo como resultados $\mathcal{X}(3)$. Esta imagen esta basada en Wang *et al.* (2018).

donde $x_n^l(i)$ es el n-th elemento de la señal de bajas frecuencias en el i-th nivel, y $\mathbf{x}^l(0)$ se establece como la señal de entrada. Las sub-señales de frecuencias bajas y altas, $\mathbf{x}^l(i)$ y $\mathbf{x}^h(i)$ en el nivel i son generados desde las 1/2 muestras paso bajo de las señales de variables intermedias definidas como (2.9) y (2.10).

$$\mathbf{a}^l(i) = \left\{a_1^l(i), a_2^l(i), \dots\right\}, \tag{2.9}$$

$$\mathbf{a}^h(i) = \left\{a_1^h(i), a_2^h(i), \dots\right\}. \tag{2.10}$$

El conjunto de sub-señales:

$$\mathcal{X}(i) = \left\{\mathbf{x}^h(1), \mathbf{x}^h(2), \dots, \mathbf{x}^h(i), \mathbf{x}^l(i)\right\}. \tag{2.11}$$

$\mathcal{X}(i)$ es llamado el i-th nivel descompuesto de $\mathbf{x}$, y este tiene diferentes resoluciones en tiempo y frecuencia. Las sub-señales con frecuencias diferentes en $\mathcal{X}$ son definidas como el MDWD, y este mantiene la misma información de orden con la señal original $\mathbf{x}$, y la frecuencia presentes desde $\mathbf{x}^h(1)$ hasta $\mathbf{x}^l(i)$ están ordenadas de altas a bajas.

La Figura 2.5 muestra una recreación de MDWD de $\mathbf{x}$ con tres niveles, cada linea punteada demarca cada nivel, la primera convolución con la señal inicial esta considerada como nivel cero. Cada rectángulo representa los filtros tanto paso bajo como paso alto dando como resultado $\mathbf{a}^l(i)$ y $\mathbf{a}^h(i)$, mientras $\mathbf{x}^l(i)$ representa la entrada de la señal para el siguiente nivel i y $\mathbf{x}^h(i)$ es adherido a la solución. Los componentes redondeados por lineas punteadas en color rojo componen la descomposición de la señal original en diversos niveles, los cuales para este estudio son utilizados como frecuencias. Finalmente como resultado se obtiene $\mathcal{X}(3)$, el cual en este caso esta compuesto por cuatro sub-señales.

2.6. Affinity Propagation (AP)

El método de agrupamiento Affinity Propagation (AP) (Frey & Dueck, 2007) es un método de agrupamiento rápido usado especialmente en los casos en los cuales los números de grupos son largos. AP trabaja en base a similaridades entre pares de datos de vectores de características (utilizando una matriz S de dimensión $n \times n$ para n puntos en el Hiperplano) y simultáneamente se consideran todos los puntos como potenciales centroides de grupos (llamados exemplars según AP).

En el algoritmo de agrupación AP, existen dos importantes conceptos: la responsabilidad $R(i,k)$ y la disponibilidad $A(i,k)$ el cual representan dos mensajes indicando qué tan adecuado es un punto de datos para ser un ejemplo potencial. $R(i,k)$ es un valor acumulado que refleja qué tan bien el punto i es adecuado para ser el candidato ejemplar del punto de datos i y luego envía datos del último al primero; es decir, en comparación con otros posibles ejemplares, el punto k es el mejor ejemplar. La disponibilidad $A(i,k)$ se opone a $R(i,k)$ y refleja cuán conveniente es que el punto i elija el punto k como su ejemplar. Basado en el punto candidato a ejemplar k, el mensaje acumulado enviado al punto de datos i indica que el punto k está más calificado que otros para ser ejemplar.

La suma de los valores de $R(i,k)$ y $A(i,k)$ es la base de evaluación para determinar si el punto de datos correspondiente puede ser un candidato ejemplar o no. Una vez que se elige un punto para ser un candidato a ejemplar, esos otros puntos con una distancia más cercana se asignarán a este grupo. Al valor similar entre dos puntos de datos x_i y x_j $(i \neq j)$ generalmente se le asigna la distancia euclidiana negativa, como $S(i,j) = -||x_i - x_j||^2$. El algoritmo utiliza un valor inicial llamado *preferencia*, que indica la preferencia de que el punto de datos se pueda elegir como ejemplar. Por lo general, se establece mediante la(s) mediana(s) de todas las distancias. El siguiente algoritmo 1 resume el proceso:

Algoritmo 1 Pseudocódigo de Affinity Propagation.

procedure ClusteringAP(S)
  $R(i,k) = 0,\ A(i,j) = 0,\ \forall i,k$
  while Until converge **do**
    $R(i,k) = S(i,k) - max(A(i,j) + S(i,j)) \mid (j \in [1,n]; j \neq k)$
    $A(i,k) = min(0, R(k,k) + \sum_j max(0, R(j,k))),\ \mid (j \in [1,n];\ j \neq i;\ j \neq k)$
    $A(k,k) = \sum_i max(0, R(i,k)),\ \mid (i \neq k)$
  end while
  return $Trks$
end procedure

El algoritmo itera hasta que los límites del clúster o grupo permanecen sin cambios durante varias iteraciones o después de una cantidad predeterminada de iteraciones. Los ejemplares (exemplars) se extraen de las matrices finales como aquellos cuya "responsabilidad + disponibilidad" para sí mismos es positiva.

2.7. K vecinos más próximos ó K-Nearest Neighbor (K-NN)

K-Nearest Neighbor (kNN) es un algoritmo de clasificación, este fue desarrollado por Evelyn Fix y Joseph Hodges en 1951. Para entender como funciona kNN primero debemos entender lo siguiente. Suponga que se tiene un conjunto de datos que contiene n atributos, los cuales podemos combinar y generar un vector n-dimensional:

$$\mathbf{x} = (x_1, x_2, \ldots, x_n)$$

Esta variable $\mathbf{x}$ también se le denomina la variable independiente. Cada ejemplo del conjunto de datos también tiene asociado otro atributo denominado y, que se denominara como variable dependiente, ya que su valor depende de los otros n atributos de $\mathbf{x}$. Se asume que y es una variable categórica, y que existe una función f, la cual asigna a la clase de la forma $y = f(\mathbf{x})$ para cada vector de los de datos de estudio. No se sabe nada de f, excepto que es una función suave de alguna forma. Se asume que se tiene un conjunto de T vectores con sus respectivas clases, es decir :

$$x^{(i)}, y^{(i)} \forall i = 1, 2, \ldots, T$$

Este conjunto es denominado conjunto de entrenamiento. El problema que se quiere resolver es, que dado un ejemplo $\mathbf{u}$, queremos determinar a que clase pertenece ese nuevo ejemplo. Si sabríamos el valor de f, simplemente bastaria con computar $v = f(\mathbf{u})$ para saber como clasificar este nuevo ejemplo, pero por supuesto no se sabe el valor de f a excepción que es una función lo suficientemente suave. La idea general de kNN es utilizar la información de los ejemplos etiquetados más similares al ejemplo dado a clasificar, utilizando para ello una distancia definida, pudiendo ser esta la distancia Ecuclidiana o la distancia de Manhattan. La idea de kNN es identificar a los k ejemplos en el conjunto de datos de entrenamiento cuyas variables independientes $\mathbf{x}$ son similares a $\mathbf{u}$, y utilizar estos k ejemplos para clasificar estos nuevos ejemplos dentro de una clase v. En el campo del aprendizaje automático es muy común utilizar el concepto de similaridad utilizando la distancia dentro del hiperplano de características.

El algoritmo kNN es uno de los algoritmos de clasificación más conocidos y un ejemplo de aprendizaje supervisado (Berastegui & Galar, 2018). Los algoritmos de vecinos cercanos (k-NN) son métodos ampliamente empleados en la clasificación estadística. Los cuales destacan por ser precisos y por no depender de ningún supuesto distribucional. A pesar de estas ventajas tienen el inconveniente de implicar un alto costo computacional. Conseguir formas eficientes de implementarlos es un reto importante para el desarrollo del reconocimiento de patrones (Villar-Patiño & Cuevas-Covarrubias, 2016). Sus propiedades teóricas garantizan que su probabilidad de error, está acotada por el doble de la probabilidad del error Bayesiano (Garcia *et al.*, 2012). En su forma más simple es un método de aprendizaje basado en casos, que conserva todos los datos de entrenamiento para clasificar, por lo que se le describe como un método del tipo "lazy learning". Sus tres limitantes más importantes en la implementación son (Villar-Patiño & Cuevas-Covarrubias, 2016):

1. Requiere espacio de almacenamiento grande para la base de entrenamiento a partir de la cual se crea la regla de decisión.
2. Bajo rendimiento en la ejecución de la regla de decisión por calcular la medida de similaridad constantemente entre las bases de entrenamiento y prueba.
3. Baja tolerancia al ruido, especialmente en la regla del vecino cercano (i.e. cuando k = 1), al considerar todos los datos como relevantes.

Una técnica que enfrenta estos tres retos es la reducción de datos, también conocida como métodos de selección de instancias o selección de prototipos. Su objetivo, es obtener un conjunto de entrenamiento representativo de tamaño mucho menor que el original y con capacidad de predicción para nuevas instancias. En otras palabras, se tiene un conjunto T, compuesto de $M + N$ instancias $x_i = (x_{i1}, x_{i2}, ..., x_{id})$, que se divide en dos conjuntos: uno de entrenamiento (BE), compuesto por M instancias y uno de prueba (BP), compuesto por N instancias, se aplica un algoritmo de Selección a la BE para generar un conjunto llamado base condensada (BC), donde BC $\subset$ BE, para clasificar a los elementos x_j de BP usando la regla kNN. Una técnica que enfrenta estos tres retos es la reducción de datos, también conocida como métodos de selección de instancias o selección de prototipos. Su objetivo, es obtener un conjunto de entrenamiento representativo de tamaño mucho menor que el original y con capacidad de predicción para nuevas instancias. En otras palabras, se tiene un conjunto T, compuesto de $M+N$ instancias $x_i = x_{i1}, x_{i2}, ..., x_{id}$, que se divide en dos conjuntos: uno de entrenamiento (BE), compuesto por M instancias y uno de prueba (BP), compuesto por N instancias, se aplica un algoritmo de Selección a la BE para generar un conjunto llamado base condensada (BC), donde BC $\subset$ BE, para clasificar a los elementos x_j de BP usando la regla kNN (Villar-Patiño & Cuevas-Covarrubias, 2016; Garcia *et al.*, 2012).

2.7.1. Ventajas y Desventajas del Algoritmo k-NN

Según el estudio de RAMOS (2016) se tiene la siguiente lista.
Ventajas:

- El coste del aprendizaje es nulo.
- No se necesita hacer suposición alguna sobre los conceptos a aprender.
- Se puede aprender conceptos complejos usando funciones sencillas como aproximaciones locales.
- Es muy tolerante al ruido.
- Es adecuado para la tarea de clasificación donde la relación entre las características son complejas y difícil de entender.

Desventajas:

- El coste de encontrar los k mejores vecinos es grande.
- No hay un mecanismo para decidir el valor óptimo para k, esto va a depende de cada conjunto de datos.
- Su rendimiento baja si el número de descriptores crece.

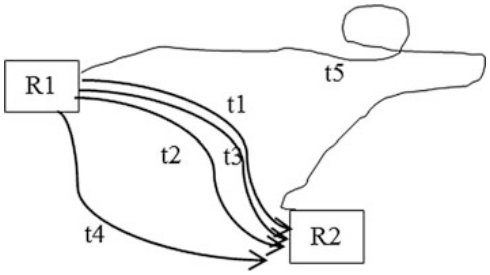
2.8. T-Distributed Stochastic Neighbor Embedding (t-SNE)

T-Distributed Stochastic Neighbor Embedding (t-Distributed Stochastic Neighbor Embedding (t-SNE)) es una técnica que nos permite visualizar datos de alta dimensión, reduciendo aquellas dimensiones en las cuales los datos se repiten. Es importante en muchos dominios diferentes y se ocupa de datos de gran dimensionalidad. Por ejemplo, los vectores de intensidad de píxeles utilizados para representar imágenes o los vectores de recuento de palabras utilizados para representar documentos, ya que normalmente estas características tienen miles de dimensiones. Los métodos de reducción de dimensionalidad convierten el conjunto de datos de alta dimensión $X = \{x_1, x_2, ..., x_n\}$ en datos bidimensionales o tridimensionales $Y = \{y_1, y_2, ..., y_n\}$ que se puede mostrar en un diagrama de dispersión. Nos referimos a la representación de datos de baja dimensión γ como un mapa, y las representaciones de baja dimensión y_i de puntos individuales como datos de puntos de mapa.

La reducción de dimensionalidad tiene como objetivo preservar la mayor parte posible de la estructura importante de los datos de alta dimensión en el mapa de baja dimensión. Las diferencias entre varias técnicas de reducción de dimensionalidad se centran en lo que preservan. Las técnicas tradicionales de reducción de dimensionalidad como Principal component analysis (PCA) (Hotelling, 1933) o técnicas clásicas de escalado multidimensional (Torgerson, 1952) son técnicas lineales que se enfocan en mantener las representaciones de baja dimensión de diferentes puntos de datos muy separados.

Para datos de alta dimensión que se encuentran en o cerca de una variedad no lineal de baja dimensión, generalmente es más importante mantener juntas las representaciones de baja dimensión de puntos de datos muy similares; esto generalmente no es posible con el mapeo lineal.

El t-SNE comienza convirtiendo las distancias euclidianas de alta dimensión entre puntos de datos en probabilidades condicionales que representan similitudes. La similitud del punto de datos x_j con el punto de datos x_i es la probabilidad condicional, $p_{j|i}$, que x_i escogería x_j como su vecino, si los vecinos se eligieran en proporción a su densidad de probabilidad bajo una distribución Gaussiana centrada en x_i. Para puntos de datos cercanos, $p_{j|i}$ es relativamente alto, mientras que para puntos de datos muy separados, $p_{j|i}$ será casi infinitesimal (para valores razonables de la varianza del Gaussiana, σ_i). Matemáticamente, la probabilidad condicional $p_{j|i}$ viene dada por:



(a)

Figura 2.6: Ejemplo de trayectoria anómala. Imagen extraída de Bhowmick & Narvekar (2018)

$$p_{j|i} = \frac{exp(-\|x_i - x_j\|^2/2\sigma_i^2)}{\sum_{k \neq i} exp(-\|x_i - x_j\|^2/2\sigma_i^2)} \tag{2.12}$$

donde σ_i es la varianza del Gaussiano que se centra en el punto de datos x_i. El método para determinar el valor de σ_i se presenta más adelante en esta sección. Debido a que solo nos interesa modelar similitudes por pares, establecemos el valor de $p_{i|i}$ en cero. Para las contrapartes de baja dimensión x_i y x_i de los puntos de datos de alta dimensión x_i y x_i, es posible calcular una probabilidad condicional similar, que denotamos por $q_{j|i}$. Para las contrapartes de baja dimensión x_i y x_i de los puntos de datos de alta dimensión x_i y x_i, es posible calcular una probabilidad condicional similar, que denotamos por $q_{j|i}$.

2.9. Trayectoria Anómala

La detección de trayectorias anómalas es todavía una área a investigar a profundidad y es claramente demandada debido a sus diferentes aplicaciones.

Definition 2.9.1. *Una trayectoria anómala es un item que es significativamente diferente a un conjunto mayoritario, esta diferencia esta definida por una forma de similaridad.*

Una trayectoria anómala tiene diferencias notable con respecto a las demás trayectorias; por ejemplo, en la Figura 2.6, $t4$ y $t5$ son trayectorias anómalas entre las regiones $R1$ y $R2$; sin embargo, $t4$ puede considerarse como un camino alterno mientras que $t5$ es en definitiva una trayectoria anómala. Las trayectorias anómalas detectadas pueden ser usadas para identificar eventos haciendo una comparación con patrones similares que fueron vistos recientemente cuando un evento irregular ocurrió (Bhowmick & Narvekar, 2018).

Existen diferentes técnicas de detección de valores atípicos en trayectorias, algunas basadas en distancias y otras basadas en densidad. Existen diferentes técnicas para la

detección de estos valores atípicos y cada uno de estos tratan de resolver éste problema con diferentes objetivos.

2.10. Medidas de Similaridad

La distancia entre dos trayectorias es usualmente medida por un tipo de distancia modificada entre los puntos de una trayectoria. Varias medidas de similaridad son definidas. Estas incluyen mínimas distancias entre pares de puntos – Closest Pair Distance (CPD), suma de pares de distancia – Sum of Pair Distance (SPD), distancia Euclidiana – Euclidean Distance (ED), Dynamic Time Warping (DTW), medidas de distancias editadas (ERP y EDR), y Longest Common Sub-Sequence (LCSS). Las medidas CPD, SPD y ED requieren de igual longitud de trayectorias y calcular la mínima distancia entre dos trayectorias; por otro lado, distancias editadas, DTW y LCSS, pueden procesar longitudes variables de trayectorias. Es importante recalcar que considerando las longitudes diferentes, bajo muestreo de puntos, incertidumbres y ruido, es difícil identificar una medida adecuada para comparar trayectorias. Todas las medidas descritas en esta sección son sensitivas al decrementos de la tasa de muestreo, ya que es un problema desafiante el procesar datos con una baja tasa de muestreo (Kim & Mahmassani, 2015b).

2.11. Demarcado (Map Matching)

Map-matching o demarcado, es una técnica de pre-procesamiento en el tratamiento de trayectorias usado para convertir una secuencia de puntos GPS a una secuencia de segmentos de tramo. Existen varias categorías de algoritmos de map-maching dependiendo del tipo de información considerada y el rango del muestreo u obtención de puntos. La categorización con el tipo de información incluyen geométrica, topológica, basadas en peso, probabilista y avanzada, donde el muestreo de puntos tiene dos tipos de métodos, el local o incremental y el global (Kim & Mahmassani, 2015b).

La mayoría de los algoritmo de demarcado asumen un muestreo de datos constante es decir con frecuencias de toma de datos alta; es por este motivo que usualmente los abordajes para este problema son locales o incrementales. También, estos abordajes consideran sólo datos espaciales para hacer un demarcado para el tramo de una determinada red, en este caso de trayectorias. Sin embargo, existe mucha información disponible que puede mejorar el demarcado de manera considerable como el tiempo, velocidad y dirección. La complejidad de los caminos que conforman el sistema urbano es otro tópico. La topología de los tramos de la red de transporte en áreas urbanas tiene muchos desafíos como caminos paralelos, caminos con múltiples capas, intercambios complejos y rutas elevadas. Todos estos factores aumenta la complejidad para procesar el demarcado, ya que éste debe ser realista y preciso.

2.12. Longitud de Trayectoria

Para nuestro estudio se hace referencia a longitud como el número de puntos que tiene una trayectoria, cada trayectoria tiene diferentes longitudes; es decir que el número de puntos que cada trayectoria conforma es diferente. Para ser más precisos, sea $T_1 = (p_1, p_2, ..., p_n)$ y $T_2 = (q_1, q_2, ..., q_m)$ trayectorias donde n y m pueden ser diferentes. Existen estudios dentro de la literatura que sólo procesan trayectorias con un número fijo de puntos, teniendo así una motivación para mejorar esta característica. Otra de las características a resaltar a la hora de describir trayectorias es la identificación de regiones para considerar el empiezo y el final de una trayectoria.

2.13. Tasas de Muestreo

Los datos GPS de vehículos en movimiento usualmente tienen una tasa baja en frecuencia para evitar la sobrecarga en costo de comunicación, espacio en disco y la duración de la batería en dispositivos móviles. El tiempo de intervalo entre dos puntos consecutivos GPS pueden ser muy largos, por ejemplo 2 minutos. Esto nos lleva a la incertidumbre que existe en la brechas de tiempo entre puntos. También la tasa de muestreo entre las trayectorias puede ser diferente; es decir, una trayectoria puede contener puntos que fueron obtenidos cada un minuto mientas que otra trayectoria en el mismo conjunto de datos puede estar conformada por puntos que son obtenidos cada dos minutos.

2.14. Dirección y Regiones

En esta sección se describen los casos en los cuales las trayectorias se consideran diferentes. Las trayectorias moviéndose en diferentes direcciones y en diferentes regiones deben ser consideradas diferentes; es decir, trayectorias próximas que están moviéndose en diferentes direcciones deben ser consideradas diferentes así como las trayectorias que se mueven en la misma dirección pero en diferentes regiones.

Como resultado del agrupamiento las trayectorias diferentes deben ser agrupadas en conjuntos diferentes. Dado que las trayectorias generadas por dispositivos GPS están relacionadas con puntos geográficos y en consecuencia estos resultados deben coincidir con un mapa geográfico (Kim & Mahmassani, 2015b).

2.15. El termino “performance” dentro de la tesis

En términos generales, la performance se refiere a qué tan bien un sistema o una organización está funcionando en relación con las expectativas y los estándares esta-

blecidos. Por ejemplo, en el contexto de un sitio web, puede ser medida en términos de velocidad de carga, tiempos de respuesta, cantidad de usuarios simultáneos que puede manejar, entre otros aspectos. De manera similar, en una organización, la performance puede ser medida en términos de productividad, eficiencia, satisfacción del cliente y otros indicadores clave de desempeño.

En el contexto de la tecnología y los negocios, el término performance se refiere a la capacidad de un sistema o una organización para cumplir con los objetivos establecidos y producir resultados efectivos y eficientes. En el mundo empresarial, la performance es un factor crítico para el éxito y la rentabilidad. Por otro lado, en el contexto científico o académico, el término performance se refiere a la capacidad de un sistema, un modelo o un método para producir resultados precisos, reproducibles y confiables en un experimento o estudio científico.

En el presente trabajo de tesis, se utiliza la exactitud (accuracy) como métrica para hacer comparaciones, así como también un análisis de eficiencia en tiempo de ejecución y consumo de memoria para dos modelos de descripción de trayectorias; es decir, dos descriptores. **La performance, para el presente trabajo de investigación, se refiere a qué tan buenos puntajes se consiguen alcanzar utilizando estas tres mediciones**.

Capítulo 3

Desarrollo del trabajo de investigación

En esta sección, se describe detalles acerca de los procedimientos que conforman nuestro modelo. Uno de los principales paso de nuestra metodología es la descripción de trayectorias; para nuestro abordaje, nos enfocamos en describir trayectorias basadas en su morfología, utilizamos esta información para agruparlas. La Figura 3.1 ilustra cada uno de los pasos aplicados en nuestro abordaje. Se divide nuestra metodología en tres módulos: Pre-procesamiento, Modelamiento y Detección.

La fase de pre-procesamiento empieza desde obtener las trayectorias, donde algunas veces es necesario utilizar algoritmos de transformación de trayectorias o metodos de limpiado de datos. Debido a que una vez obtenidas las trayectorias desde el mundo real, estas presentan ruido, además que los datos no se encuentran estandarizados. El modelado de trayectorias consiste en buscar una representación adecuada, esta representación subraya características que ayudan a discriminar o clasificar nuestros datos para un propósito especifico.

La extracción de características esta presente en este paso. Finalmente, en el tercer módulo, la detección de anomalías, propiamente este paso consiste en detectar un punto aislado en un hiperplano para detectar anomalías.

3.1. Datos de trayectorias

El modulo de Datos consiste en la obtención y el uso de trayectorias en nuestro estudio. Para nuestro abordaje, una trayectoria señaliza un objeto en movimiento. Teniendo esta como extremos el inicio y el fin del tramo. En este trabajo, se utiliza tres conjuntos de datos sintéticos y dos conjuntos de datos reales.

- Los datos sintéticos creados en Piciarelli *et al.* (2008), esta conformado por $260,000$ trayectorias generadas por un algoritmo.Estas trayectorias fueron divididas en $1,000$ grupos, y cada grupo contiene 260 instancias con coordenadas

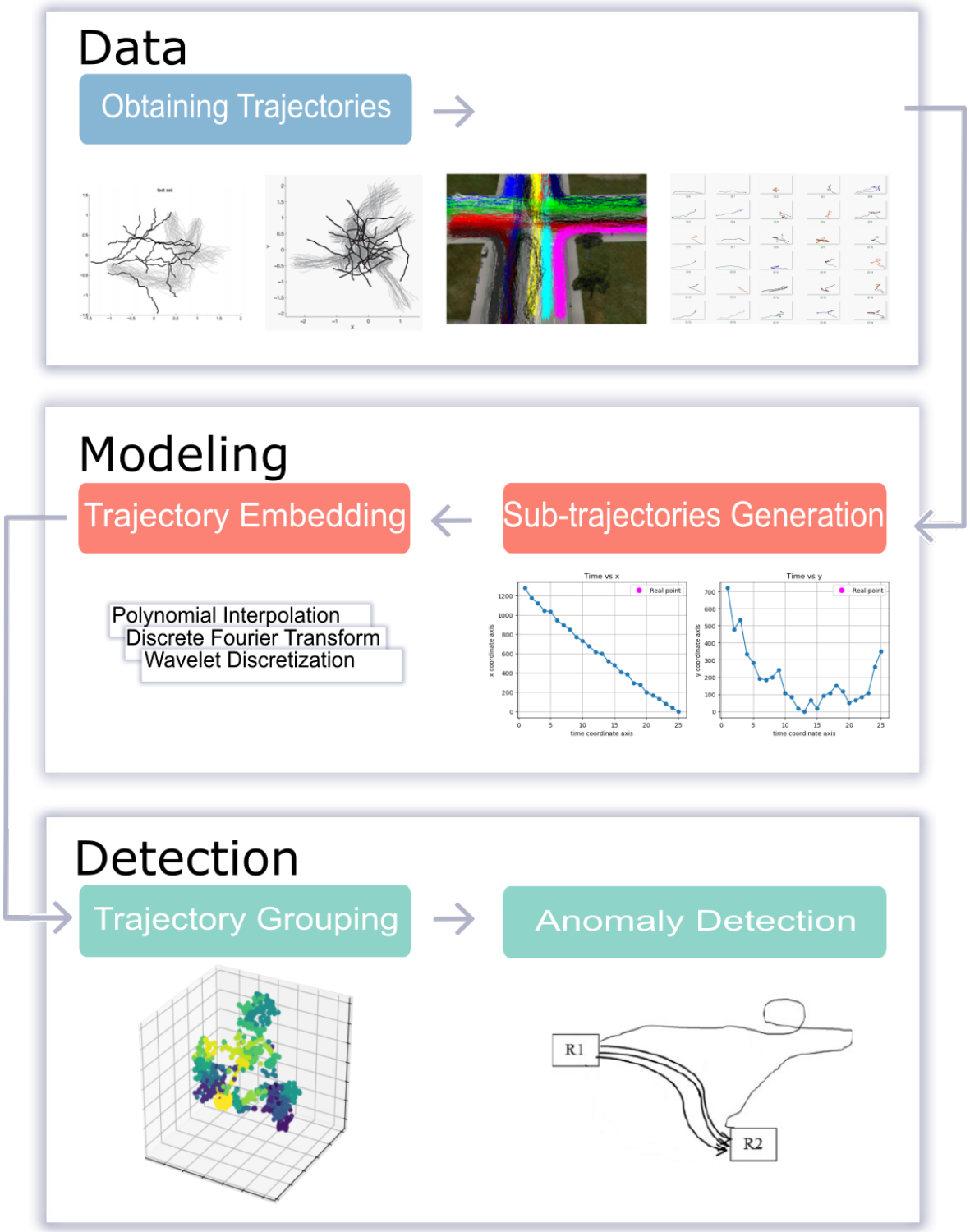


Figura 3.1: Descripción general del modelo propuesto, basado en tres módulos: Datos, Modelado y detección de anomalías.

(x, y), de cuyos 250 pertenecen a cinco grupos, y las diez ultimas trayectorias son anómalas. Estas trayectorias tienen 16 puntos de longitud, sin información de tiempo.

- El conjunto de datos fue creado por Laxhammar & Falkman (2014) usando el algoritmo de Piciarelli *et al.* (2008). Este conjunto de datos es nuevo y esta conformado por 200, 000 trayectorias, con 100 grupos pertenecientes a 10 diferentes grupos, cada grupo contiene 2, 000 trayectorias. Este conjunto de datos permiten realizar pruebas en eficiencia y en efectividad.
- El conjunto de datos CROSS (Morris & Trivedi, 2011) contiene 9, 700 trayectorias simulando una intersección de tráfico de cuatro pistas o caminos, con varios patrones de giros o rutas, incluyendo giros en forma de U. El conjunto de datos consiste de 9, 500 actividades representadas por caminos, pertenecientes a 19 grupos y 200 caminos anómalos. Estos caminos tienen diferentes longitudes sin información de tiempo.
- Un conjunto de datos denominado ***Laboratory*** con 1, 000 trayectorias, estas trayectorias con coordenadas (x, y). Este conjunto contiene 970 trayectorias normales y 30 anómalas. Estos recorridos tienen diferentes longitudes sin información de tiempo. Estas trayectorias fueron extraídas desde vídeos que pertenecen a un laboratorio. Los vídeos fueron usados para analizar anomalías, eventos en situaciones simples (Mora *et al.*, 2020), el contenido es real sin forzar ninguna situación anormal.
- Un conjunto de datos denominado ***Traffic-Flow***, que consta de 1,107,795 registros, que luego fueron transformados en 15,013 trayectorias. Estos datos corresponden a un intervalo de tiempo continuo. Los recorridos tienen diferentes longitudes con información de tiempo en cada punto de cada trayectoria. Estas trayectorias fueron extraídas desde dispositivos GPS que son utilizados para hacer monitoreo a los ómnibus, que pertenecen a empresas de transporte público, el contenido es real sin forzar ninguna situación anormal.

Todos estos conjuntos de datos son normalizados y limpiados; por lo tanto, ningún tipo de pre-procesamiento fue aplicado sobre estos datos. En esta sección, es importante mencionar que nuestro abordaje soporta trayectorias de diferentes longitudes; esto se corrobora con la experimentación de los conjuntos de datos pertenecientes a Laboratory y CROSS, que contienen trayectorias con diferentes longitudes.

3.2. Modelamiento de Trayectorias

A continuación, procederemos a describir el modelamiento de trayectorias en detalle. Desde una vez obtenido los datos hasta la generación de vectores de características.

3.2.1. *Normalización de Trayectoria*

Este proceso fue aplicado a cada trayectoria. Para este propósito utilizamos el método de *feature scaling*.

Se definen los siguientes espacios:

$$w_x = \{x_i \in p_i \mid \forall p_i \in T_j\}, \tag{3.1}$$

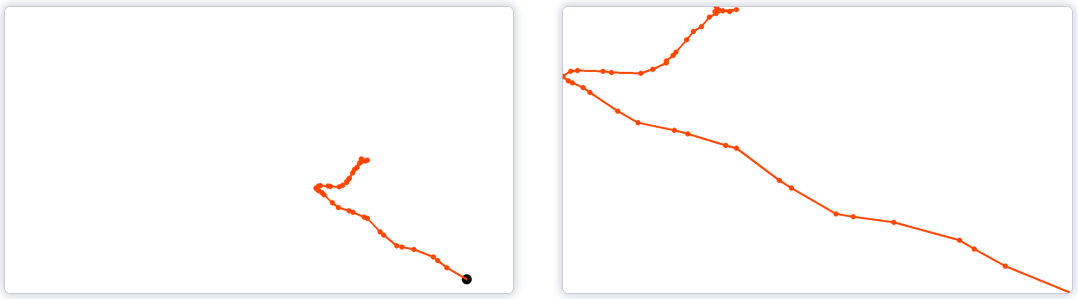
$$w_y = \{y_i \in p_i \mid \forall p_i \in T_j\}, \tag{3.2}$$

Donde p_i es un punto y T_j una trayectoria. Para todas las variables x_i y y_i de T_j, las fórmulas del método *feature-scaling* son aplicadas:

$$x'_i = \frac{x_i - min(w_x)}{max(w_x) - min(w_x)}, \tag{3.3}$$

$$y'_i = \frac{y_i - min(w_y)}{max(w_y) - min(w_y)}, \tag{3.4}$$

donde *min* y *max* retornan los valores mínimo y el máximo de un espacio respectivamente. Después de computar cada componente w_x y w_y con (3.3) y (3.4), cada elemento tiene un nuevo valor asignado. Se asigna el valor de cero para el mínimo, y uno para el máximo, y el resto de valores intermedios son escalas entre esos limites. Por ejemplo, para propósitos de visualización, se multiplica como valor maximo la dimension de 720 y 1280 por cada componente respectivamente (x and y), obteniendo; como resultado, la figura 3.2. b (Fig. 3.2.a muestran la trayectoria original).



(a) (b)

Figura 3.2: La normalización de trayectorias mejoran las características. (a) Esta trayectoria fue recogida de un segmento de vídeo de vigilancia, haciendo un seguimiento a una persona. (b) En esta imagen cada componente de la trayectoria a sido normalizado hacia las dimensiones del vídeo original, mostrando así visualmente que efecto causa la normalización en este caso.

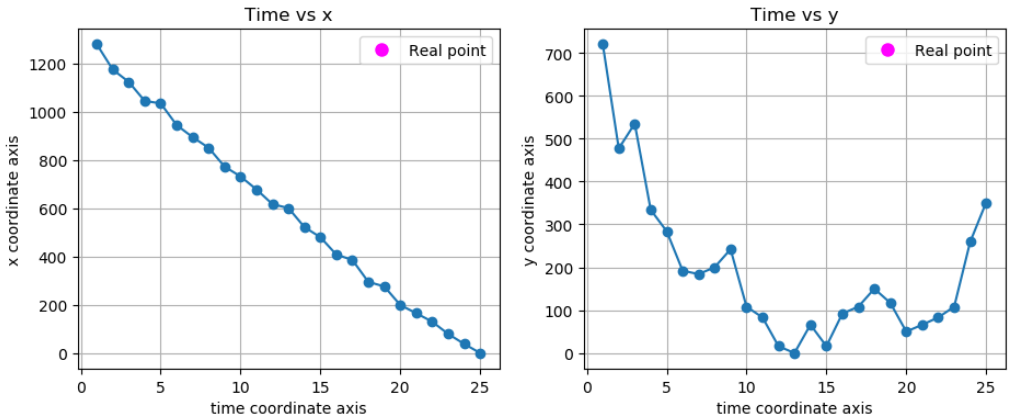


Figura 3.3: Descomposición de Trayectoria. Nuestro modelo separa una trayectoria en dos conjuntos de puntos unidimensional (1D).

3.2.2. *Descomposición de Trayectoria*

Para esta subsección, la representación de trayectorias serán generadas por la partición de ellas en sub-trayectorias 1-D para los espacio x e y, representado como $X = x_i, Y = y_i, i = 1, \ldots, n$ (n es el numero de puntos de una trayectoria), X y Y representan los movimientos horizontal y vertical respectivamente. Fig. 3.3 muestra un ejemplo de dos sub-trayectorias generadas por nuestro modelo. Otra interpretación que encaja dentro de este proceso es la del comportamiento de cada sub-trayectoria como ondas o series temporales, representando cada una de estas, la variación de cada componente espacial. Por lo tanto, las señales dan un comportamiento parametrizado en comparación con los datos que normalmente te da una trayectoria, esta información ayuda a la descripción de forma de una trayectoria.

3.2.3. *Representación de Espacio de Características*

Una vez finalizada la *descomposición de trayectorias*, es posible aplicar los métodos de extracción de características para describir cada sub-trayectoria. Las entradas para el descriptor son las dos sub-trayectorias. Se considera dos técnicas para alcanzar la descripción de las sub-trayectorias, que incluyen a las transformadas discretas de Fourier y Wavelet. La derivación y la representación de nuestro espacio de características de las sub-trayectorias usando los tres métodos propuestos es especificado como sigue:

Transformada Discreta de Fourier La representación del espacio de características usando DFT es de la siguiente manera. Los N-puntos producidos por DFT de X (revisar 2), define como una secuencia X_f de N números complejos ($f = 0, ..., N-1$), y esta dado por (3.5), y tambien en forma similar para Y puede definirse por (3.6).

$$X_f = DFT(X) \tag{3.5}$$

$$Y_f = DFT(Y) \tag{3.6}$$

X_f e Y_f son números complejos con la excepciona de X_0, Y_0 los cuales son reales. Como una regla, la secuencia de DFT esta truncada después de m términos para X_f y k para Y_f. Formalmente, sean a_i y $\hat{a}_i$ las partes real e imaginaria de X_f, y sean b_i y $\hat{b}_i$ las partes real e imaginaria de Y_f. Debido a que se define trabajar con números reales en vez de números imaginarios, se convierte X_f e Y_f en números reales, usando (3.7) y (3.8) respectivamente.

$$r_i = \sqrt{{a_i}^2 + \hat{a}_i^2}, i = 0, ..., m-1 \tag{3.7}$$

$$\bar{r}_j = \sqrt{{b_i}^2 + \hat{b}_i^2}, j = 0, ..., k-1 \tag{3.8}$$

con los números r_i y $\bar{r}_j$, notamos que en su mayoría de ellos aparecen dos veces, escogemos sólo uno de estos números como en (3.9) y (3.10).

$$R_x = \{r_0, ..., r_i, ..., r_{m-1}\}_{\neq} \tag{3.9}$$

$$R_y = \{\bar{r}_0, ..., \bar{r}_j, ..., \bar{r}_{k-1}\}_{\neq} \tag{3.10}$$

donde R_x y R_y son conjuntos formados por elementos únicos. Se ejecuta la discretización o *binning* con estos dos conjuntos de variables, transformando la longitud variable de un conjunto de variables a uno de longitud constante. Esto fuer realizado utilizando histogramas b_q y b'_q. Estos están sometidos a las siguientes condiciones:

$$|R_x| = \sum_{q=1}^{l} b_q \tag{3.11}$$

$$|R_y| = \sum_{q=1}^{l} b'_q \tag{3.12}$$

donde b_q y b'_q son funciones para contar el numero de observaciones que encaja cada categoría disconjunta (bins). l es el numero de bins, $|R_x|$ y $|R_y|$ son los números de observaciones de R_x y R_y respectivamente. Finalmente, la trayectoria puede ser representada en el espacio de caracteristicas por $\mathbf{F}_{DFT}$ definido en (3.13).

$$\mathbf{F}_{DFT} = [\sum_{q=1}^{l} b_q, \sum_{q=1}^{l} b'_q] \tag{3.13}$$

Descomposición Discreta Multinivel de Wavelet Para el proceso de descomposición con MDWD, se utiliza la familia de Wavelets del tipo *Haar*. Esta técnica representa diferentes niveles de frecuencias dependiendo de las diferentes formas presentadas en la trayectoria. Aplicando MDWD en X y Y, es posible obtener (3.14) y (3.15) respectivamente.

$$[cA_m, cD_m, cD_{m-1}, \ldots, cD_2, cD_1] = MDWD(X) \tag{3.14}$$

$$[cA_k, cD_k, cD_{k-1}, \ldots, cD_2, cD_1] = MDWD(Y) \tag{3.15}$$

La salida es una lista de coeficientes, donde m y k denota el nivel máximo útil de descomposición. Por lo tanto, el primer elemento cA_m del resultado es el arreglo aproximado de coeficientes, y los siguientes elementos $cD_m, \ldots, cD_1$ son arreglos mas detallados de coeficientes.

Se define el vector de características F_x como la concatenación de diferentes niveles de coeficientes obtenidos con MDWD para X esta dado por (3.16). Una expresión similar puede definirse para Y como (3.17).

$$\mathbf{F}_x = [cA_m, cD_m, cD_{m-1}, \ldots, cD_2, cD_1] \tag{3.16}$$

$$\mathbf{F}_y = [cA_k, cD_k, cD_{k-1}, \ldots, cD_2, cD_1] \tag{3.17}$$

con F_x y F_y, se ejecuta la discretización usando histogramas b_q y b'_q, estas cumplen las siguientes condiciones:

$$|F_x| = \sum_{q=1}^{l} b_q \tag{3.18}$$

$$|F_y| = \sum_{q=1}^{l} b'_q \tag{3.19}$$

De forma similar como fue aplicado con DFT. Finalmente, la trayectoria puede ser representada en el espacio de características por $\mathbf{F}$ definido como (3.20).

$$\mathbf{F} = [m, k, \sum_{q=1}^{l} b_q, \sum_{q=1}^{l} b'_q] \tag{3.20}$$

3.3. Detección de Anomalía

Una vez obtenidos los vectores de características, se procede a ejecutar la detección de anomalías. Para este objetivo se utilizará el abordaje denominado *distance-based* mencionado por Zhang *et al.* (2020). Este enfoque no dice que aquellas trayectorias que se encuentran a una larga distancia desde otro conjunto de trayectorias son recogidas como anómalas; es necesario que exista un conjunto agrupado de trayectorias y otras cuantas de manera isolada y distantes, para que este concepto se aplique. Para el propósito de detección de estas anomalias se utiliza agrupación automática (aprendizaje no supervisado).

Se segmenta y se separa la información de trayectorias en el proceso de agrupación automática para detectar anomalías (los cuales están localizados en los extremos lejos de los grupos mayoritarios). Como método de agrupación automática, se utiliza AP, este método se adapta a nuestro abordaje. Además, AP permite la separación de diferentes trayectorias debido a que este método de agrupamiento genera mayor numero de grupos en comparación con otros métodos de aprendizaje no supervisado, para este caso en particular. Finalmente, para recuperar las trayectorias anómalas desde los grupos, se define un umbral. El umbral de anomalia es definido como el numero máximo de elementos que debe tener un grupo anómalo.

Capítulo 4

Pruebas y resultados

En esta sección se describirá el resultado de nuestros experimentos. Tres conjuntos de datos sintéticos fueron utilizados, ademas de dos conjunto de datos reales denominado *Laboratory* y *Traffic Flow*. Estos conjuntos de datos nos permiten obtener resultados tanto cualitativos como cuantitativos.

4.1. Configuración Experimental

Los hiperparámetros para nuestros experimentos son descritos de la siguiente manera. Para el algoritmo de agrupamiento AP, en cuanto al parámetro denominado *preference parameter* es configurado hacia la mediana de similaridad de las entradas, y el factor denominado *damping factor* es configurado con los valores 0,5; 0,625 y 0,7. En algunos casos, el máximo numero de iteraciones debe ser configurado a mil. En el caso de los histogramas, el numero de bins (barras del histograma) es configurado a diez. Para la obtención de los valores del *average accuracy* (exactitud media), se escoge el mejor umbral que detecta a las anomalías (*Anomaly threshold*), esto en cada subconjunto de trayectorias con las que se experimentó.

4.2. Evaluación

Para evaluar el rendimiento relativo de nuestros descriptores en un conjunto de datos exhaustivo, se ejecutan los experimentos utilizando los datos sintéticos generados por Piciarelli *et al.* (2008), Laxhammar & Falkman (2014). Debido a la naturaleza y el tamaño de estos conjuntos de datos, se utilizará la métrica denominada *precisión media* (average-accuracy) para evaluar el rendimiento. Sin embargo, en los conjuntos de datos denominado CROSS y *Laboratory*, se utiliza sólo precisión, debido a que éstos están compuestos de un sólo conjunto de trayectorias. Los resultados son presentados en la Tabla 4.1.

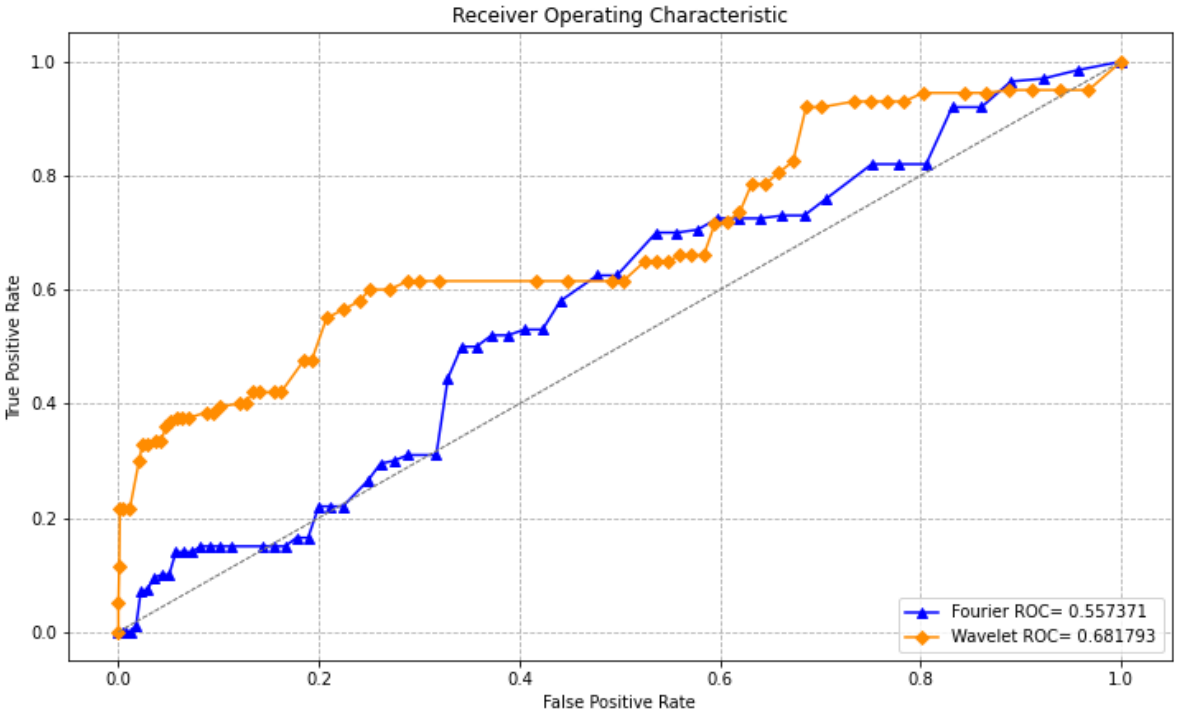


Figura 4.1: Evaluación comparativa de la performance de MDWD y DFT utilizando el conjunto de datos CROSS.

Además para la evaluación de CROSS y *Laboratory*, se decide evaluar el performance de estos conjuntos de datos utilizando la curva ROC, debido a que existen diferentes umbrales para detectar anomalías, y también debido a los trabajos relacionados de Mora *et al.* (2020) y Dias *et al.* (2020). Cada punto de la curva ROC es obtenido con un valor diferente de umbral anómalo (Anomaly-Threshold), denotando los valores de TPR y FPR en cada detección de anomalía. Esta información provee una percepción visual del mejor umbral usando FPR y TPR. La Figura 4.1 muestra el resultado obtenido para el conjunto de datos CROSS. De acuerdo a esta figura, el descriptor de MDWD consigue un mejor resultado a comparación con DFT en éste conjunto de datos.

Cuadro 4.1: Resultados cuantitativos para los conjuntos de datos sintéticos.

	Conjuntos de Datos		
Descriptores	Piciarelli	Laxhammar	CROSS
MDWD	0.9519	0.9780	**0.8884**
DFT	**0.9525**	**0.9848**	0.8825

4.3. Discusión

Primero, se describe los resultados alcanzados con datos sintéticos, y después se explica los resultados obtenidos con el conjunto datos denominado CROSS.

Observando la Tabla 4.1 se puede decir que, para los primeros dos conjuntos de datos, esta claro que DFT tiene mayor performance en la detección. En el tercer conjunto de datos, los resultados para DFT fueron menores en comparación con MDWD. A pesar de la menor precisión obtenida con ambos descriptores en el conjunto de datos CROSS. Este conjunto de datos presenta mayor dificultad en el su procesamiento y descripción de su comportamiento, debido también a que la longitud de sus trayectorias es variable.

Observando nuevamente la Tabla 4.1, la mejor puntuación obtenida es con el conjunto de datos de *Laxhammar*, y se considera que este resultado es competitivo con aquellos trabajos presentes en la literatura como son Piciarelli *et al.* (2008) y Ergezer & Leblebicioğlu (2016). Por otro lado, para comparar nuestros resultados obtenidos con la base datos CROSS, utilizaremos los resultados obtenidos por Morris & Trivedi (2011), quien utilizó la técnica de Naftel & Khalid (2006) y quien identificó 84 % de anomalías con un 10 % de tasa de falsos positivos. Para nuestros caso usando TPR se obtiene 64 % con 24 % de tasa de falsos positivos. Obteniendo resultados prometedores, debido a que nuestra representación toma en cuenta la información relativa a la morfología de una trayectoria y la base de datos CROSS recolecta información de forma en su definición de anomalía, se considera que este conjunto de datos recolecta información similar a los objetivos propuestos en este estudio, por consiguiente estos resultados tendrían mayor relevancia en comparación con los dos primeros conjuntos de datos mencionados.

Capítulo 5

Caso de estudio

5.1. Detección de ruta anómala

El segundo caso de estudio conlleva a identificar trayectorias anómalas en datos generados por buses de transporte público. La idea es detectar cuando un autobús de transporte se sale de su ruta o cuando los dispositivos utilizados están defectuosos. Para lograrlo, utilizamos el conjunto de datos generados en el trabajo de Alvarez Mamani (2018), los datos producto de este trabajo son generados por dispositivos GPS. Este trabajo consistió en desarrollar un *Intelligent Transportation System (ITS)* para monitorear buses de transporte urbano.

5.1.1. Base de Datos

El conjunto de datos Traffic FLow contiene diferentes rutas de unidades de transporte público. Considera las diferencias entre los ciclos que el bus completa durante el día. Un ciclo incluye las rutas de ida y vuelta. El servicio de monitoreo de buses es utilizado por más de una empresa de transporte. El proceso de monitoreo es constante, los dispositivos GPS funcionan las veinticuatro horas y los puntos de datos se producen cada 30 segundos. Para el estudio, el criterio para definir una trayectoria "normal" viene dado por la ruta determinada por la empresa de transporte público elegida. Por otro lado, los criterios para definir trayectoria anómala son: la ruta que la empresa de transporte no define, o la ruta mal generada por los errores del dispositivo GPS, por su mala instalación o desgaste.

Este conjunto de datos ofrece una gran cantidad de datos de monitoreo, por lo que se necesito realizar un proceso de selección. Los criterios de selección de datos son los siguientes: elegimos una empresa de transporte, para este caso de estudio se selecciona la **empresa de transporte Arcoíris**, la cual tiene como ruta la universidad, el hecho de elegir una empresa de transporte a la vez también ayuda a reducir la enorme cantidad de datos a analizar. Elegimos como fecha de estudio, del 02 de mayo al 06 de junio de

2019 debido a que este período corresponde al último mes del conjunto de datos que poseemos. Se selecciona el horario entre cinco y veintidós horas para considerar sólo el horario de trabajo de los buses.

La selección realizada produjo un conjunto de datos que consta de 1,107,795 registros, transformados en 15,013 trayectorias. Para construir las trayectorias, se consideraron los siguientes cuatro atributos proporcionados por la base de datos: el *día*, la *placa* del autobús, el *número de viaje* y la *hora* en que se crearon los registros. El atributo de *número de viaje* permite indicar cuándo el bus logró un ciclo, incluyendo la ida y la vuelta en un tramo de viaje. Con los tres primeros atributos mencionados, creamos subconjuntos de registros, para después con el atributo de tiempo, ordenarlos en cada subconjunto. Para construir las trayectorias se utilizó el orden cronológico de cada registro. Es necesario agrupar estos atributos ya que cualquiera de estos buses puede realizar múltiples viajes en un solo día, y es necesario distinguirlos.

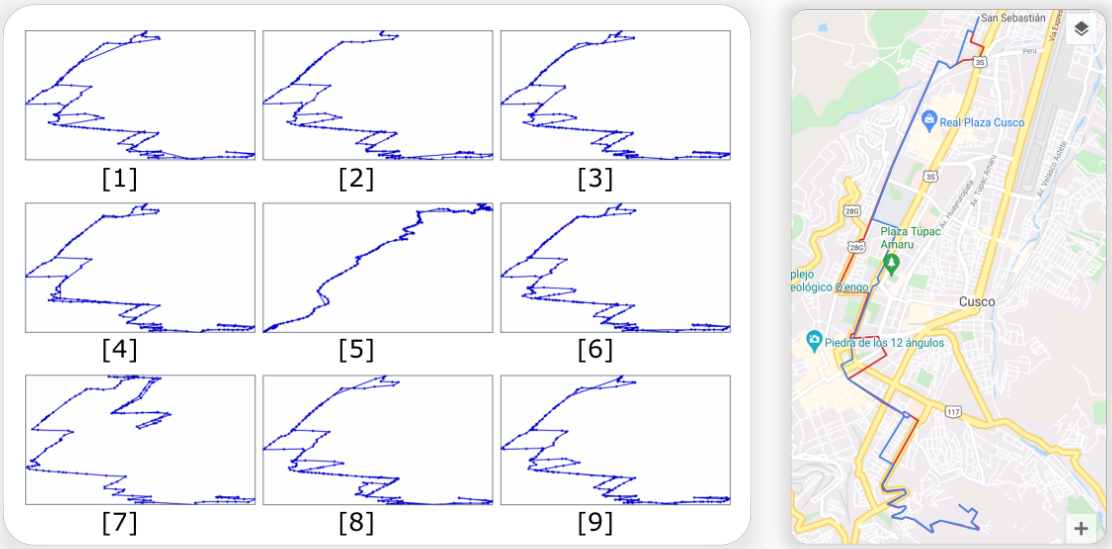
Finalmente, una vez que construimos las trayectorias, fue necesario eliminar los puntos ruidosos generados por algunos errores de los dispositivos GPS. Para este propósito, logramos dos pasos: consideramos solo los puntos cuya diferencia en el plano de latitud y longitud es menor a una unidad y no consideramos trayectorias con menos de veinte puntos. Los criterios son los siguientes: a) Esta diferencia de unidades se tomó debido a que los buses no pueden moverse más de una unidad de medida en esos planos en tan poco tiempo, y b) una ruta no puede definirse por tan pocos puntos (veinte) en esta base de datos generada.

5.1.2. Extracción de características y agrupamiento

Para los próximos pasos, elegimos solo MDWD como descriptor, ya que para este caso de estudio, la idea no es comparar dos descriptores, sino mostrar los resultados de nuestro modelo utilizando datos de una aplicación real. Elegimos MDWD teniendo en cuenta los resultados previos obtenidos en el primer caso de estudio.

Para este segundo caso de estudio dividimos los datos en cuatro grupos, esto debido a la gran cantidad de datos que se tiene para procesar. Los cuatro conjuntos de datos tienen la siguiente distribución: el primer conjunto de trayectorias consta de $3,754$ elementos, y los siguientes tres conjuntos de trayectoria, constan de $3,753$ de elementos cada uno. El proceso de agrupamiento se aplica una vez ya generada los vectores de características para cada conjunto creado. Con estos conjuntos de datos, generamos 93, 71, 70 y 78 grupos respectivamente.

Como umbral de anomalía, consideramos los grupos que contienen uno y dos elementos. La Fig. 5.1.a muestra trayectorias normales, a excepción de dos (Fig. 5.1.a.5 y Fig. 5.1.a.7), que son detectadas como anómalas. Según la Fig. 5.1.a, es posible notar en esta, que estas trayectorias anómalas son diferentes al grupo mayoritario que son definidas como normales. La Fig. 5.1.b muestra la ruta seleccionada para este caso de estudio. Cada una de las rutas está dibujada y disponible en la aplicación del sistema inteligente realizado por Alvarez Mamani (2018). Este dibujo realizado en Google Maps



(a) (b)

Figura 5.1: Detecciones de trayectorias anómalas en el conjunto de datos *Traffic Flow*. Fig. (a.5) y Fig. (a.7) son detectadas como anómalas por nuestro abordaje. Mientras que en (b) se muestra la ruta seleccionada vista desde Google Maps, esto es generado en el aplicativo de la empresa.

nos permite visualizar dónde se encuentra la ruta y distingue las rutas de ida y vuelta con diferentes colores.

5.1.3. Descubrimientos

Este caso de estudio tiene como objetivo mostrar una herramienta para visualizar las rutas anómalas o errores constantes en los dispositivos GPS utilizados. Por ello, presentamos algunos resultados en cuanto a la evaluación cualitativa de nuestro abordaje. Con la ayuda de los administradores del ITS y Google Maps, fue posible comparar y validar nuestros resultados.

Primero, se encontró la fecha en la cual se produjeron las trayectorias anómalas, esto con los atributos de *fecha* y *hora*, que nos proporciona la base de datos. La Fig. 5.1.a.5 muestra a la trayectoria anómala producida el 19 de mayo desde las 5 horas hasta las 21 horas de un domingo. Preguntando a los administradores del ITS, encontramos que esta trayectoria es diferente debido a que la empresa, los días domingos trabaja de manera irregular, y estos días no se respeta estrictamente la ruta trazada por la empresa. En cuanto a la segunda trayectoria anómala (Fig. 5.1.a.7), ésta se produjo un jueves 09 de mayo de 07 horas a 21 horas. Al buscar la ruta en Google Maps y preguntar a los administradores del ITS, concluimos que en esta fecha, debido a las festividades, se presentó un desfile en la ciudad de Cusco, lo que le produjo al conductor la necesidad de utilizar rutas alternas para terminar su viaje en el tiempo establecido

por la empresa.

Capítulo 6

Eficiencia en Tiempo y Memoria

Para evaluar el rendimiento de los descriptores, medimos el tiempo de ejecución y el consumo de memoria. Nuestros experimentos se realizaron con un procesador Intel(R) Core(TM) i7-4710HQ CPU @ 2.50 GHz 2.50 GHz con 8 GB de RAM y Windows 10 Pro x64 bits. El pre-procesamiento de la trayectoria (normalización) se realizó utilizando el lenguaje C++, mientras que el detector de trayectorias anómalas utiliza las siguientes bibliotecas de Python: *Scipy* y *Numpy* (Walt *et al.*, 2011), *Scikit-learn* (Pedregosa *et al.*, 2011), *PyWavelets* (Lee *et al.*, 2019) y el proceso de medir la eficiencia fue realizada con las librerías *Time* y *Memory-Profile* de Python.

Para cada caso de estudio, se muestrea diferentes subconjuntos. La división de los conjuntos de datos se realizó teniendo en cuenta la variación de tamaño, para estas mediciones de menor a mayor. La distribución de los conjuntos de datos es la siguiente: se crearon tres conjuntos de trayectorias por cada caso de estudio, y cada uno de estos conjuntos tiene un número diferente de trayectorias como se mencionó. Para el primer caso de estudio, el número de elementos de cada conjunto es cien, mil y dos mil. Para el segundo caso de estudio, el número de elementos de cada conjunto es mil, dos mil y quince mil.

Medimos el tiempo de ejecución y la memoria consumida por cada subconjunto. Se obtuvieron valores promedio para cada sub-conjunto: para el caso del tiempo, se promediaron siete mediciones para obtener un valor promedio de tiempo de ejecución y, en el caso de la memoria, se promediaron tres mediciones para obtener un valor de consumo de memoria promedio. La Fig. 6.1 muestra cada uno de estos valores promedio para cada tamaño de conjunto de trayectoria diferente.

Cabe resaltar que sólo se consideró la generación de características para las medidas de tiempo y memoria. La lectura, la escritura y la transformación de variables de cadena a números no se considera al medir el tiempo y la memoria de los descriptores, ya que estos procesos pueden alterar los resultados de medición. Cada descriptor termina procesando $2N$ señales por cada N-trayectorias, debido a que cada trayectoria se divide en dos sub- trayectorias.

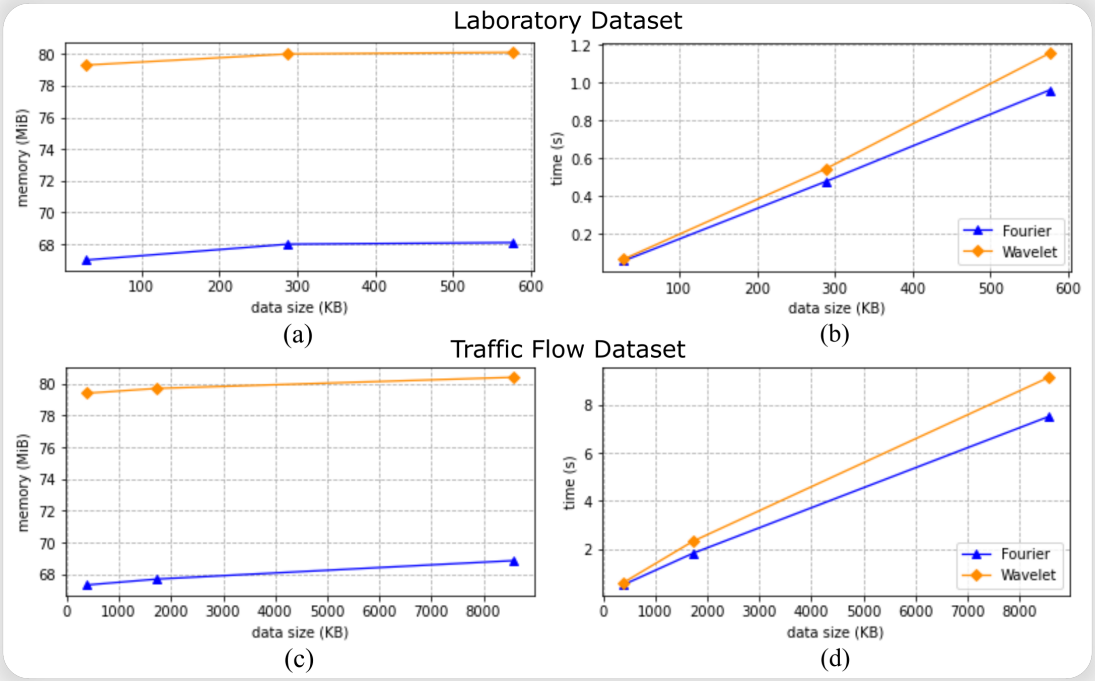


Figura 6.1: Gráficos de los valores promedio de memoria y de tiempo de ejecución. Estos valores dependen de los pesos de cada conjunto de datos de entrada. Estos valores se obtuvieron por cada caso de estudio.

Para mostrar la escalabilidad bajo diferentes tamaños de los datos de entrada, graficamos nuestros resultados con diferentes tamaños de datos, estos medidos en *kilobytes* (KB). La Figura 6.1.a y la Fig. 6.1.c muestran la medida de eficiencia de la memoria en *mebibyte* (MiB). Por otra parte la Fig. 6.1 .b y la Fig. 6.1.d muestran la medida de eficiencia del tiempo de ejecución en *segundos* (s).

Debido al número variable de puntos de las trayectorias en ambos casos de estudio, decidimos graficar estos datos en función de los pesos de cada archivo. De acuerdo con la Fig. 6.1, podemos observar que mientras crece el tamaño de los datos de entrada, la memoria utilizada no escala tanto como el tiempo de ejecución. Debido al pequeño consumo de memoria y al corto tiempo de ejecución, podemos obtener resultados en poco tiempo.

La complejidad de nuestro enfoque está limitada por la complejidad algorítmica del método de agrupamiento AP ya que este algoritmo tiene una complejidad algorítmica de orden $O(N^2T)$ según Refianti *et al.* (2017), donde N es el número de muestras y T es el número de iteraciones hasta la convergencia el algoritmo. El algoritmo de AP tiene un orden de $O(N^2)$ en términos de complejidad de memoria. Al mismo tiempo, los algoritmos MDWD y DFT tienen un orden de complejidad $O(NlogN)$ tanto en el tiempo y memoria según Wickerhauser (1996). Estos ordenes de complejidad son confirmados por las medidas obtenidas en esta sección.

En esta sección mostramos que es posible obtener resultados en tiempo real para detectar trayectorias anómalas, incluso con un hardware limitado. Debido a la complejidad de estos algoritmos, nuestra implementación tiene una ventaja en velocidad y almacenamiento. Adicionalmente, de acuerdo con la literatura, el agrupamiento de AP se puede mejorar en tiempo de ejecución con la utilización de tarjetas gráficas. Según Shi (2017) es posible mejorar el desempeño del algoritmo de agrupamiento usando Graphics Processing Unit (GPU).

La hipótesis planteada afirmaba que la performance del descriptor de Wavelet sería mejor que la del descriptor de Fourier en la detección de trayectorias anómalas. Tras realizar los experimentos y analizar los resultados, podemos concluir que esta hipótesis ha sido confirmada.

Como se vio en esta sección se llevaron a cabo experimentos de rendimiento, midiendo el tiempo de ejecución y el consumo de memoria de ambos descriptores. Los resultados revelaron que ambos descriptores se comportaron de manera similar en cuanto a rendimiento; es decir, no se encontraron diferencias significativas en el tiempo de ejecución ni en el consumo de memoria entre el descriptor de Wavelet y el descriptor de Fourier.

Sin embargo, es importante destacar que la métrica de precisión y la curva ROC fueron los elementos diferenciadores clave en la comparación de rendimiento. Estos resultados respaldan la afirmación de que el descriptor de Wavelet tiene un desempeño superior al de Fourier en términos de precisión en la detección de trayectorias anómalas. Estos hallazgos son importantes para el desarrollo de técnicas de detección de anomalías en aplicaciones que involucran análisis de trayectorias.

Capítulo 7

Conclusiones y Trabajos Futuros

7.1. Conclusiones

1. Este trabajo presenta un análisis comparativo de dos descriptores de trayectorias usando una representación de espacio basada en coeficientes, generando características para detectar anomalías en trayectorias. En este trabajo también se esta introduciendo MDWD como un extractor de características que considera la morfología de las trayectorias, y éste arroja resultados satisfactorios comparados con otros descriptores, aportando así en el campo de la detección de trayectorias anómalas, debido al desempeño previsto por este método.

2. Se verifica el rendimiento del métodos de aprendizaje no supervisados denominado AP en la detección de trayectorias anómalas. Se observa que el número de clases o tipos diferentes de trayectorias en el conjunto de datos utilizado, influye en el rendimiento del algoritmo de agrupamiento; es decir, si el número de clases de trayectorias incrementa, la precisión del agrupamiento decrementará. Esto se pudo recoger gracias a las observaciones hechas a los experimentos con diferentes algoritmos de clustering como Density-based spatial clustering of applications with noise (DBSCAN) Además nuestro estudio aporta al campo de análisis de similaridad de trayectorias.

3. Se demuestra la usabilidad de nuestro abordaje en la detección de trayectorias anómalas, utilizando bases de datos tanto sintéticos como reales. Las bases de datos utilizadas se encuentran presentes en la literatura y de dominio público, así como también bases de datos recientemente creados. La utilidad de nuestro abordaje a sido demostrado a través de los experimentos para detectar anomalías en el conjunto de datos denominado *Laboratory* y *Traffic Flow*, que representan acontecimientos reales.

Debido a que la performance del descriptor de Wavelet, específicamente MDWD, presenta mejor rendimiento frente al descriptor de Fourier en la detección de trayectorias anómalas basadas en su morfología, esto utilizando la métrica conocida como

curva ROC, la hipótesis de investigación (H_i) es apoyada por los datos obtenidos en este estudio, aportando evidencia a favor de la misma.

7.2. Trabajos futuros

Existe la posibilidad de mejorar el proceso de aprendizaje no supervisado usando el método denominado *Adaptive AP* (Wang *et al.*, 2008), para seleccionar de manera automática el parámetro denominado *preference-parameter* y encontrar el agrupamiento optimo para la solución. También el uso de *k-Nearest Neighbor* para no tener que definir el umbral anómalo de manera manual, permitiendo así la detección automática de este umbral que permite separar las características anómalas, esto último dentro del espacio de características en el hiperplano.

Bibliografía

Alvarez Mamani, Edwin. 2018. Sistema de transporte inteligente (STI), para el control y monitoreo del servicio urbano en la Ciudad del Cusco. Universidad Nacional de San Antonio Abad del Cusco.

Annoni, Ronald, & Forster, Carlos HQ. 2012. Analysis of aircraft trajectories using fourier descriptors and kernel density estimation. *Pages 1441–1446 of: 2012 15th International IEEE Conference on Intelligent Transportation Systems*. IEEE.

Berastegui, AG, & Galar, IM. 2018. Implementación del algoritmo de los k vecinos más cercanos (k-NN) y estimación del mejor valor local de k para su cálculo. *Trabajo Fin de Grado). Recuperado de: https://academicae. unavarra. es/bitstream/handle/2454/29112/Memoria. pdf*.

Bhowmick, Kiran, & Narvekar, Meera. 2018. Trajectory outlier detection for traffic events: A survey. *Pages 37–46 of: Intelligent Computing and Information and Communication*. Springer.

Cooley, JW, Lewis, P, & Welch, P. 1969. The finite Fourier transform. *IEEE Transactions on audio and electroacoustics*, **17**(2), 77–85.

Dee, Hannah M, & Hogg, David C. 2004. Detecting inexplicable behaviour. *Pages 1–10 of: BMVC*. Citeseer.

Dias, Madson LD, Mattos, César Lincoln C, da Silva, Ticiana LC, de Macedo, José Antônio F, & Silva, Wellington CP. 2020. Anomaly detection in trajectory data with normalizing flows. *Pages 1–8 of: 2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE.

Ergezer, Hamza, & Leblebicioğlu, Kemal. 2016. Anomaly detection and activity perception using covariance descriptor for trajectories. *Pages 728–742 of: European Conference on Computer Vision*. Springer.

Frey, Brendan J, & Dueck, Delbert. 2007. Clustering by passing messages between data points. *science*, **315**(5814), 972–976.

Garcia, Salvador, Derrac, Joaquin, Cano, Jose, & Herrera, Francisco. 2012. Prototype selection for nearest neighbor classification: Taxonomy and empirical study. *IEEE transactions on pattern analysis and machine intelligence*, **34**(3), 417–435.

Guo, Diansheng, Liu, Shufan, & Jin, Hai. 2010. A graph-based approach to vehicle trajectory analysis. *Journal of Location Based Services*, **4**(3-4), 183–199.

Han, Binh, Liu, Ling, & Omiecinski, Edward. 2012. Neat: Road network aware trajectory clustering. *Pages 142–151 of: 2012 IEEE 32nd International Conference on Distributed Computing Systems*. IEEE.

Hernández-Sampieri, Roberto, & Torres, Christian Paulina Mendoza. 2018. *Metodología de la investigación*. Vol. 4. McGraw-Hill Interamericana México.

Hotelling, Harold. 1933. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, **24**(6), 417.

Khalid, Shehzad, & Naftel, Andrew. 2010. Automatic motion learning in the presence of anomalies using coefficient feature space representation of trajectories. *Acta Automatica Sinica*, **36**(5), 655–666.

Kim, Jiwon, & Mahmassani, Hani S. 2015a. Spatial and temporal characterization of travel patterns in a traffic network using vehicle trajectories. *Transportation Research Procedia*, **9**, 164–184.

Kim, Jiwon, & Mahmassani, Hani S. 2015b. Spatial and temporal characterization of travel patterns in a traffic network using vehicle trajectories. *Transportation Research Procedia*, **9**, 164–184.

Kong, Xiangjie, Li, Menglin, Ma, Kai, Tian, Kaiqi, Wang, Mengyuan, Ning, Zhaolong, & Xia, Feng. 2018. Big trajectory data: A survey of applications and services. *IEEE Access*, **6**, 58295–58306.

Labs, INRIA. 2004. *CAVIAR "INRIA" Dataset*.

Laxhammar, Rikard, & Falkman, Göran. 2014. Online learning and sequential anomaly detection in trajectories. *IEEE transactions on pattern analysis and machine intelligence*, **36**(6), 1158–1173.

Lee, Gregory, Gommers, Ralf, Waselewski, Filip, Wohlfahrt, Kai, & O'Leary, Aaron. 2019. PyWavelets: A Python package for wavelet analysis. *Journal of Open Source Software*, **4**(36), 1237. doi:10.21105/joss.01237.

Lee, Jae-Gil, Han, Jiawei, & Whang, Kyu-Young. 2007. Trajectory clustering: a partition-and-group framework. *Pages 593–604 of: Proceedings of the 2007 ACM SIGMOD international conference on Management of data*.

Mora, Rensso, Cayllahua, Edward, de Melo, Victor C., Cámara-Chávez, Guillermo, & R. Schwartz, William. 2020. Anomaly Event Detection based on People Trajectories for Surveillance Videos. *Pages 107–116 of: VISIGRAPP - 15th International Conference on Computer Vision Theory and Applications*.

Morris, Brendan Tran, & Trivedi, Mohan Manubhai. 2011. Trajectory learning for activity understanding: Unsupervised, multilevel, and long-term adaptive approach. *IEEE transactions on pattern analysis and machine intelligence*, **33**(11), 2287–2301.

Moumena, Ahmed. 2016. Anomalies Detection Based on the ROC Analysis using Classifiers in Tactical Cognitive Radio Systems: A survey. *IAES International Journal of Artificial Intelligence (IJ-AI)*, **5**(3), 105–116. doi: 10.11591/ij-ai.v4i3.1415.

Naftel, Andrew, & Khalid, Shehzad. 2006. Classifying spatiotemporal object trajectories using unsupervised learning in the coefficient feature space. *Multimedia Systems*, **12**(3), 227–238.

Panagiotakis, Costas, Pelekis, Nikos, Kopanakis, Ioannis, Ramasso, Emmanuel, & Theodoridis, Yannis. 2011. Segmentation and sampling of moving object trajectories based on representativeness. *IEEE Transactions on Knowledge and Data Engineering*, **24**(7), 1328–1343.

Pang, Yutian, & Liu, Yongming. 2020. Conditional Generative Adversarial Networks (CGAN) for Aircraft Trajectory Prediction considering weather effects. *Page 1853 of: AIAA Scitech 2020 Forum*.

Parent, Christine, Spaccapietra, Stefano, Renso, Chiara, Andrienko, Gennady, Andrienko, Natalia, Bogorny, Vania, Damiani, Maria Luisa, Gkoulalas-Divanis, Aris, Macedo, Jose, Pelekis, Nikos, *et al.* 2013. Semantic trajectories modeling and analysis. *ACM Computing Surveys (CSUR)*, **45**(4), 1–32.

Paul, Saswata, Hole, Frederick, Zytek, Alexandra, & Varela, Carlos A. 2017. Flight trajectory planning for fixed-wing aircraft in loss of thrust emergencies. *arXiv preprint arXiv:1711.00716*.

Pedregosa, Fabian, Varoquaux, Gaël, Gramfort, Alexandre, Michel, Vincent, Thirion, Bertrand, Grisel, Olivier, Blondel, Mathieu, Prettenhofer, Peter, Weiss, Ron, Dubourg, Vincent, *et al.* 2011. Scikit-learn: Machine learning in Python. *the Journal of machine Learning research*, **12**, 2825–2830.

Piciarelli, Claudio, Micheloni, Christian, & Foresti, Gian Luca. 2008. Trajectory-based anomalous event detection. *IEEE Transactions on Circuits and Systems for video Technology*, **18**(11), 1544–1554.

Powell, Mark D, & Aberson, Sim D. 2001. Accuracy of United States tropical cyclone landfall forecasts in the Atlantic basin (1976–2000). *Bulletin of the American Meteorological Society*, **82**(12), 2749–2768.

Quispe Torres, Gerar Francis, Colque, Rensso Victor Hugo Mora, & Schwartz, William Robson. 2019. Surveillance video summarization based on trajectory rarity measure. Universidad Católica San Pablo.

RAMOS, FELIPE ALBERTO CIFUENTES. 2016. *Clasificación automática de Tweets utilizando K-NN y K-Means como algoritmos de clasificación automática, aplicando TF-IDF y TF-RFL para las ponderaciones*. Ph.D. thesis, Pontificia Universidad Católica de Valparaíso.

Refianti, R, Mutiara, AB, & Gunawan, S. 2017. Time complexity comparison between affinity propagation algorithms. *Journal of Theoretical & Applied Information Technology*, **95**(7).

Rinzivillo, Salvatore, Pedreschi, Dino, Nanni, Mirco, Giannotti, Fosca, Andrienko, Natalia, & Andrienko, Gennady. 2008. Visually driven analysis of movement data by progressive clustering. *Information Visualization*, **7**(3-4), 225–239.

Roh, Gook-Pil, & Hwang, Seung-won. 2010. Nncluster: An efficient clustering algorithm for road network trajectories. *Pages 47–61 of: International Conference on Database Systems for Advanced Applications*. Springer.

Scott, David W. 2015. *Multivariate density estimation: theory, practice, and visualization*. John Wiley & Sons.

Shi, Xuan. 2017. Parallelizing affinity propagation using graphics processing units for spatial cluster analysis over big geospatial data. *Pages 355–369 of: Advances in Geocomputation*. Springer.

Sillito, Rowland R, & Fisher, Robert B. 2008. Semi-supervised Learning for Anomalous Trajectory Detection. *Pages 035–1 of: BMVC*, vol. 1.

Smith, Steven W, *et al.* 1997. The scientist and engineer's guide to digital signal processing. Chapter 8: The Discrete Fourier Transform. ISBN:978-0-9660176-3-2.

Torgerson, Warren S. 1952. Multidimensional scaling: I. Theory and method. *Psychometrika*, **17**(4), 401–419.

Turchini, Francesco, Seidenari, Lorenzo, & Del Bimbo, Alberto. 2015. Understanding sport activities from correspondences of clustered trajectories. *Pages 43–50 of: Proceedings of the IEEE International Conference on Computer Vision Workshops*.

Villar-Patiño, Carmen, & Cuevas-Covarrubias, Carlos. 2016. Controlled condensation in k-NN and its application for real time color identification. *Revista de Matemática Teoría y Aplicaciones*, **23**(1), 143–154.

Walt, Stéfan van der, Colbert, S Chris, & Varoquaux, Gael. 2011. The NumPy array: a structure for efficient numerical computation. *Computing in Science & Engineering*, **13**(2), 22–30. doi:10.1109/MCSE.2011.37.

Wang, Jingyuan, Wang, Ze, Li, Jianfeng, & Wu, Junjie. 2018. Multilevel wavelet decomposition network for interpretable time series analysis. *Pages 2437–2446 of: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*.

Wang, Kaijun, Zhang, Junying, Li, Dan, Zhang, Xinna, & Guo, Tao. 2008. Adaptive affinity propagation clustering. *arXiv preprint arXiv:0805.1096*.

Wang, Xiaogang, Ma, Keng Teck, Ng, Gee-Wah, & Grimson, W Eric L. 2011. Trajectory analysis and semantic region modeling using nonparametric hierarchical bayesian models. *International journal of computer vision*, **95**(3), 287–312.

Wickerhauser, Mladen Victor. 1996. *Adapted wavelet analysis: from theory to software*. AK Peters/CRC Press.

Wisdom, Michael J, Cimon, Norman J, Johnson, Bruce K, Garton, Edward O, & Thomas, Jack Ward. 2004. Spatial partitioning by mule deer and elk in relation to traffic. *In: In: Transactions of the 69th North American Wildlife and Natural Resources Conference: 509-530.*

Xu, Hongteng, Zhou, Yang, Lin, Weiyao, & Zha, Hongyuan. 2015. Unsupervised trajectory clustering via adaptive multi-kernel-based shrinkage. *Pages 4328–4336 of: Proceedings of the IEEE International Conference on Computer Vision.*

Zhang, Haiyan, Luo, Yonglong, Yu, Qingying, Sun, Liping, Li, Xuejing, & Sun, Zhenqiang. 2020. A framework of abnormal behavior detection and classification based on big trajectory data for mobile networks. *Security and Communication Networks*, **2020**.