Introducción a la Topología Computacional

La topología computacional es una rama interdisciplinaria de las matemáticas y la informática que busca comprender y analizar la forma y la estructura de los conjuntos de datos, especialmente aquellos que provienen de contextos espaciales y geométricos. Esta disciplina utiliza herramientas matemáticas y algoritmos computacionales para revelar patrones subyacentes, propiedades topológicas y características geométricas en datos complejos.

La cantidad de datos recaudada por instituciones públicas y privadas ha explotado en los últimos 15 años gracias a la creciente cobertura de las redes de internet y la disminución del costo de almacenamiento de información. En el año 2000 se almacenaron a nivel mundial 800 mil petabytes (PB) de datos y esta cantidad aumenta constantemente.

En la actualidad, por ejemplo, X (anteriormente Twitter) genera siete terabytes (7 TB) de datos diariamente, Facebook 10 TB. Se calcula que la cantidad de datos almacenada anualmente alcance 35 zettabytes (ZB=un billón de terabytes).

Se denomina Big Data a un conjunto de información tan grande, complejo y, en la mayoría de los casos, sin estructura, que resulta imposible estudiarlo con las herramientas usuales de manejo de base de datos.

El estudio del manejo del Big Data incluye retos como la optimización de la captura, almacenamiento, búsqueda, trasferencia, análisis, visualización, etc. Actualmente en muchísimas ramas de la ciencia y la industria se tiene acceso a bases de datos gigantescas con información cruda de la cual se pueden extraer patrones, relaciones y en un siguiente paso, teorías. Una de las nuevas técnicas desarrolladas es el Análisis Topológico de Datos (TDA, por sus siglas en ingles), este se ha practicado con éxito en los últimos 15 años para estudiar cómo se puede inferir información de un sistema de datos a partir de muestras representadas como un espacio topológico combinatorio.

En el TDA se construyen conjuntos asociados a los datos y se infieren características cualitativas del conjunto a partir de la homología de dicho complejo.

1. **Análisis de Imágenes Médicas:**
    - *Problema:* Identificar y cuantificar las estructuras morfológicas en imágenes médicas.
    - *Aplicación de Topología Computacional:* Utilizando complejos simpliciales para representar la conectividad entre píxeles, la topología computacional puede ayudar a identificar regiones anómalas o patrones específicos en imágenes de resonancia magnética (IRM) o tomografías computarizadas (TC).

2. **Biología Computacional - Datos Genómicos:**
    - *Problema:* Comprender la estructura y relaciones en conjuntos de datos genómicos.
    - *Aplicación de Topología Computacional:* La topología computacional puede emplearse para analizar la forma en que genes interactúan y se expresan, representando conexiones entre genes mediante complejos simpliciales. Esto permite identificar patrones topológicos relevantes para la biología molecular.

3. **Análisis de Redes Sociales:**
    - *Problema:* Comprender la estructura y conexiones en redes sociales.
    - *Aplicación de Topología Computacional:* La topología computacional puede revelar patrones de conexiones y comunidades en redes sociales mediante la representación de usuarios como nodos y relaciones como aristas. La identificación de componentes conectados y la persistencia de grupos a lo largo del tiempo son ejemplos de aplicaciones.

4. **Procesamiento de Señales en Ingeniería:**
    - *Problema:* Analizar patrones en señales complejas.
    - *Aplicación de Topología Computacional:* Representando señales como funciones y aplicando funciones de filtración, se pueden identificar características topológicas, como picos y valles, utilizando la homología persistente. Esto puede ser útil en el análisis de señales biomédicas, señales acústicas o cualquier conjunto de datos basado en el tiempo.

5. **Diseño de Circuitos Electrónicos:**
    - *Problema:* Optimizar el diseño de circuitos electrónicos complejos.
    - *Aplicación de Topología Computacional:* Representar la conectividad entre componentes electrónicos como complejos simpliciales permite analizar la

topología del diseño. La identificación de regiones críticas o conexiones redundantes puede mejorar la eficiencia del diseño.

6. **Estudio del Paisaje:**
    - *Problema:* Analizar la topología del terreno en un área geográfica.
    - *Aplicación de Topología Computacional:* Representar elevaciones geográficas como complejos simpliciales para identificar características topológicas en el paisaje. Esto puede ser útil en la planificación urbana, la gestión de recursos naturales o la evaluación de riesgos geológicos.

7. **Estudio de Redes de Transporte:**
    - *Problema:* Analizar la conectividad y eficiencia de una red de transporte.
    - *Aplicación de Topología Computacional:* Representar nodos como estaciones o intersecciones y enlaces como conexiones entre ellas. La topología computacional puede identificar patrones de congestión, áreas críticas y rutas eficientes en sistemas de transporte.

8. **Diseño de Materiales en Nanotecnología:**
    - *Problema:* Optimizar la estructura de materiales a escala nanométrica.
    - *Aplicación de Topología Computacional:* Utilizar complejos simpliciales para representar la disposición de átomos en materiales nanoestructurados. La topología computacional puede ayudar a identificar propiedades emergentes y optimizar el diseño de nuevos materiales.

9. **Análisis de Datos Climáticos:**
    - *Problema:* Estudiar patrones climáticos y eventos extremos.
    - *Aplicación de Topología Computacional:* Representar datos climáticos espaciales y temporales utilizando complejos simpliciales. La topología computacional puede ayudar a identificar patrones persistentes como tormentas, patrones de viento y cambios climáticos significativos.

10. **Reconocimiento de Patrones en Imágenes Satelitales:**
    - *Problema:* Detectar patrones geográficos y cambios en grandes conjuntos de datos de imágenes satelitales.
    - *Aplicación de Topología Computacional:* Utilizar la topología computacional para analizar la distribución de características geográficas y cambios en la vegetación a lo largo del tiempo. Esto puede ser valioso en la monitorización de la deforestación, cambios urbanos y gestión del territorio.

12. **Análisis de Datos de Redes Neuronales:**
    - *Problema:* Estudiar la conectividad y patrones de activación en redes neuronales artificiales.
    - *Aplicación de Topología Computacional:* Representar la arquitectura de una red neuronal como un complejo simplicial. La topología computacional puede revelar patrones emergentes, nodos críticos y relaciones en la red, proporcionando información sobre su funcionamiento.

Herramientas en Python :

1. **GUDHI:**
    - *Descripción:* GUDHI (Geometry Understanding in Higher Dimensions) es una biblioteca dedicada a la topología computacional. Proporciona implementaciones eficientes de algoritmos para calcular la homología persistente, así como herramientas para construir complejos simpliciales.
    - *Sitio web:* GUDHI en GitHub
2. **Ripser:**
    - *Descripción:* Ripser es una herramienta eficiente para calcular la homología persistente en conjuntos de datos. Es fácil de usar y proporciona una interfaz de línea de comandos para realizar cálculos de homología persistente de manera rápida.
    - *Sitio web:* Ripser en GitHub
3. **PHAT (Persistent Homology Algorithms Toolbox):**
    - *Descripción:* PHAT es una biblioteca que proporciona algoritmos para el cálculo de homología persistente. Ofrece implementaciones en Python y C++ y es especialmente útil para el análisis de grandes conjuntos de datos.
    - *Sitio web:* PHAT en GitHub
4. **Dionysus:**
    - *Descripción:* Dionysus es una biblioteca de topología computacional que implementa algoritmos para el cálculo de homología persistente. Ofrece una interfaz sencilla y es compatible con Python 2 y 3.
    - *Sitio web:* Dionysus en GitHub
5. **TDA (Topological Data Analysis):**
    - *Descripción:* TDA es una biblioteca que proporciona herramientas para el análisis topológico de datos, incluyendo el cálculo de homología persistente y la construcción de complejos simpliciales.
    - *Sitio web:* TDA en GitHub

Metodologías

KDD

La metodología KDD se caracteriza por identificar patrones dentro de un conjunto de datos, el proceso consta de manera general de cinco pasos:

- Selección del conjunto de datos.
- Preprocesamiento.
- Transformación.
- Minería de Datos.
- Evaluación.
- Interpretación.

Este método utiliza principalmente aprendizaje semiautomático debido al preprocesamiento que existe de los datos seleccionados previamente, lo que ayuda a la transformación de dichos datos para adaptarlos al algoritmo que será utilizado en el siguiente paso (minería de datos). La principal actividad en minería de datos es la búsqueda de patrones en el conjunto de datos a analizar. Una vez terminado el proceso de minería de datos se realiza una evaluación e interpretación de los datos, esto implica analizar los resultados obtenidos en los pasos ejecutados previamente, en caso de que los resultados no sean significativos se debe iniciar el proceso nuevamente.

Objetivos.

- Búsqueda de patrones en bases de datos.
- Implementar técnicas de clasificación para localizar patrones.

Ventajas.

- Flujo especifico para el pre procesamiento de los datos.
- Utilizado en minería de datos.
- Funciona muy bien con información estructurada.

Desventajas.

- No tiene buenos resultados en información no estructurada.
- Falta una fase que realice una delimitación específica del problema a solucionar.
- No existe una conexión entre las diferentes fases que ayude a retroalimentar si se esta siguiendo el flujo correcto. Hasta la fase final es donde se pueden visualizar resultados.

Semma

Se caracteriza principalmente por un enfoque orientado a análisis a nivel estadístico de la información procesada. Esta metodología utiliza el mayor numero de variables predictivas que son fáciles de detectar con el fin de determinar el nivel de exactitud del modelo.

Los pasos por seguir en la metodología Semma son los siguientes:

- Muestra: Obtiene un numero significativo de elementos en el conjunto a analizar con un enfoque balanceado entre el consumo de recursos computacionales.
- Explorar: Localización de patrones en la información previamente seleccionada, este proceso debe ser iterativo con el fin de localizar conjuntos de datos que ayuden a determinar si la muestra fue seleccionada de manera eficiente.
- Modificar: En caso de detectar nuevos patrones en la información en el conjunto de datos previamente seleccionados se deberá realizar las modificaciones pertinentes al algoritmo o método estadístico a utilizar en el proceso de análisis/ejecución.
- Modelo: En este paso es cuando se prueba el algoritmo utilizado, las pruebas serán orientadas a revisar que este funcionando de manera adecuada.
- Evaluación: Esta área se caracteriza por la evaluación de manera detallada de los resultados obtenidos por el modelo utilizado en el paso anterior.

Los trabajos de investigación que utilizan Semma como metodología siguen los pasos anteriormente mencionados adaptando la metodología a la problemática que estén atendiendo.

Objetivos.

- Búsqueda de patrones en bases de datos.
- Limpieza y preparación de la información.
- Implementación de un algoritmo para la localización de patrones.

Ventajas.

- Flujo especifico para el pre procesamiento de los datos.
- Permite la modificación del proceso en una fase previa antes de la obtención de los resultados finales.
- Utilizado en minería de datos.
- Funciona muy bien con información estructurada.

Desventajas.

- No tiene buenos resultados con información no estructurada.
- La fase que permite la modificación de los algoritmos o técnicas utilizadas para la limpieza de la información es en un proceso complejo y podría incrementar el tiempo para obtener resultados satisfactorios.
- No existe una interacción entre las diferentes fases que ayude a retroalimentar si se esta siguiendo el flujo correcto.

Crisp-DM

La metodología Crisp-DM fue desarrollada en el año de 1996 con el fin de obtener patrones de conjuntos de datos utilizando algoritmos del área de minería de datos. Esta metodología consta de los siguientes pasos:

- Entendimiento del negocio: Se realiza un análisis profundo sobre el problema que se va a solucionar, el principal objetivo es acotar el alcance del problema.
- Entendimiento de los datos: Una vez acotado el problema se debe seleccionar el conjunto de datos orientados al problema que se planea solucionar.
- Preparación de los datos: Se limpia la información que pueda generar ruido en el modelado, usualmente se remueven datos poco significativos.
- Modelado: Ejecución del algoritmo para obtener los patrones en el conjunto de datos previamente seleccionados.
- Evaluación: Se realizan evaluaciones a nivel estadístico para determinar la eficiencia del algoritmo ejecutado previamente.
- Ejecución: En caso de tener una evaluación satisfactoria se procede a utilizar el algoritmo en ambientes de producción.

Esta metodología posee como característica principal el análisis profundo del problema que se va a solucionar, esto ayuda a delimitar el problema resultando en un conjunto de datos acotado de manera ideal para representar una problemática especifica.

Objetivos.

- Búsqueda de patrones en bases de datos.
- Limpieza y preparación de la información.
- Mantener retroalimentación constante entre todas las fases de la metodología
- Acotar el problema.
- Implementación de un algoritmo para la localización de patrones.

Ventajas.

- Flujo específico para el preprocesamiento de los datos.
- Tiene una fase para acotar el problema a solucionar con la metodología.
- Permite la modificación del proceso entre cada una de las fases obteniendo retroalimentación del proceso por fases.
- Utilizado en minería de datos.
- Funciona muy bien con información estructurada y no estructurada.

Desventajas.

- La comunicación constante entre diferentes fases puede confundir al investigador y el sobreuso de dicha comunicación incrementa el tiempo para obtener resultados.