INSTITUTO POLITÉCNICO NACIONAL

CENTRO DE INVESTIGACIÓN EN COMPUTACIÓN

Tarea 03

Clasificación

Edgar Fernando Espinosa Torres

Clasificación inteligente de patrones

Dr. Cornelio Yáñez Márquez

CIUDAD DE MÉXICO, 16 DE SEPTIEMBRE DE 2023.

Índice general

Índice de tablas

Introducción

El reconocimiento de patrones (RP) es una rama crucial de la inteligencia artificial que se centra en la identificación y clasificación de patrones en conjuntos de datos. La literatura en este campo es vasta y proviene de múltiples disciplinas, incluyendo minería de datos, aprendizaje automático y, por supuesto, el propio reconocimiento de patrones.

La comprensión de las semejanzas y diferencias entre las metodologías y enfoques de estas disciplinas puede ofrecer una perspectiva única y enriquecedora. Así, esta tarea número tres tiene como objetivo explorar estas semejanzas y diferencias a través de la revisión de tres libros clave en el ámbito del RP y el aprendizaje automático.

Además, se lleva a cabo un análisis de un dataset específico para entender y aplicar conceptos prácticos relacionados con multi-label classification.

Desarrollo y discusión

Propósito de la tarea

Verificar que existen disciplinas íntimamente relacionadas con el RP.

Identificar datasets para multi-label classification.

Parte 1

Se analizaron los siguientes libros:

- Data Mining: Concepts and Techniques de Han, J. et al.
- Introduction to Machine Learning de Alpaydin, E.
- Pattern Recognition de Theodoridis, S y Koutroumbas, K.

A continuación, expongo algunas de las semejanzas y diferencias entre los tres libros [1] [2] [3], enfocándose en algunos de los temas:

Semejanzas:

- Preprocesamiento y conocimiento de datos:

"Data Mining": En el capítulo 2, se discuten los tipos de datos, descripciones estadísticas, visualización y medidas de similitud.

"Introduction to Machine Learning": Aunque no hay un capítulo específico, el preprocesamiento de datos es una parte esencial del aprendizaje automático y se infiere en múltiples capítulos.

"Pattern Recognition": El capítulo 5 se centra en la selección de características, que es una forma de preprocesamiento.

- Métodos paramétricos y no paramétricos:

"Data Mining": Estos conceptos se encuentran subyacentes en capítulos como clasificación y clustering.

"Introduction to Machine Learning": Discutido en los capítulos 4 (métodos paramétricos) y 8 (métodos no paramétricos).

"Pattern Recognition": Estos conceptos se reflejan en el enfoque de diferentes clasificadores y técnicas.

- Árboles de decisión:

"Data Mining": Se infiere en los capítulos de clasificación.

"Introduction to Machine Learning": Capítulo 9.

"Pattern Recognition": Incluido en los métodos de clasificadores no lineales en el capítulo 4.

- Modelos basados en Redes Neuronales:

"Data Mining": Se infiere en los capítulos de clasificación avanzada.

"Introduction to Machine Learning": Capítulo 11 sobre perceptrones multicapa y aprendizaje profundo.

"Pattern Recognition": Discutido en el capítulo 4 sobre clasificadores no lineales y redes neuronales.

- Modelos gráficos:

"Introduction to Machine Learning": Capítulo 14 sobre modelos gráficos.

"Pattern Recognition": Parte del capítulo 9 sobre clasificación dependiente del contexto.

Diferencias:

- Profundidad en temas específicos:

"Data Mining": Da un enfoque amplio sobre cómo obtener patrones y conocimientos de grandes conjuntos de datos, incluyendo temas como OLAP y tecnología de cubo de datos.

"Introduction to Machine Learning": Se centra en el desarrollo y comprensión de algoritmos para aprender patrones. Temas como máquinas de kernel y modelos ocultos de Markov son exclusivos de este libro.

"Pattern Recognition": La atención se centra en la identificación y clasificación de patrones específicos, con capítulos dedicados a la generación de características y el emparejamiento de plantillas.

- Tendencias y aplicaciones:

"Data Mining": Tiene un enfoque más aplicado, con un capítulo (13) dedicado a tendencias y fronteras de investigación.

"Introduction to Machine Learning": Aunque aborda aplicaciones prácticas, no tiene un capítulo específico dedicado a las tendencias actuales.

"Pattern Recognition": Su enfoque es más teórico y metodológico.

- Enfoque metodológico:

"Data Mining": Más centrado en la aplicación y la práctica.

"Introduction to Machine Learning": Balance entre teoría y aplicación.

"Pattern Recognition": Más teórico y basado en fundamentos.

Parte 2

Encontrar un dataset que esté diseñado para multi-label classification.

2.1 Reportar el nombre del dataset

El nombre del dataset es flags [4], el cual contiene atributos como área del país, idioma, religión, entre otros.

2.2 Reportar el repositorio de donde se obtuvo el dataset

El dataset se obtuvo de Mulan [4].

2.3 Reportar un ejemplo de patrón que esté repetido en diferentes clases y decir en cuáles clases está repetido ese patrón

1,4,0,0,1,1,0,0,6,0,1,1,1,0,0,0,1,1,1,1,1,1,1,1,1,0,1

1,4,0,0,1,1,0,0,6,0,1,1,1,0,0,0,1,1,0,1,1,1,1,1,1,0,1

Número de patrón	Clases a las que pertenece	Vector de etiquetas
25	Icon, Animate, Text, Red, Green, Blue, Yellow, White, Orange	[0,0,1,1,1,1,1,1,1,1,0,1]
177	Icon, Animate, Red, Green, Blue, Yellow, White, Orange	[0,0,1,1,0,1,1,1,1,1,0,1]

Tabla 1: Patrón repetido del dataset flag.

2.4 Explicar brevemente cómo debería modificarse ese dataset a fin de que se convierta en un dataset útil para el curso CIP.

Para ello vamos a revisar las recomendaciones dadas en las diapositivas número 03 correspondientes a la clase de CIP, las cuales fueron:

- Al trabajar por primera vez con un dataset, lo primero que se debe hacer es buscar patrones repetidos en diferentes clases. Si eso ocurre, significa que ese dataset no es para single-label classification, sino que pertenece a la rama multi-label classification, la cual está fuera del curso CIP.

Comentario: En este dataset se localizaron al menos dos patrones repetidos en diferentes clases.

- Otra posibilidad es eliminar ese patrón de todas las clases donde se repita, y así trabajar en single-label classification.

Comentario: Es fácil eliminar el único patrón repetido.

- Luego, se deben buscar rasgos que tengan los valores iguales (o perdidos) en todos los patrones. Esos rasgos se eliminan.

Comentarios: Esto no ocurre en el dataset utilizado.

- A continuación, se deben buscar patrones cuyos valores en todos sus rasgos sean perdidos. Esos patrones se eliminan.

Comentarios: No hay valores perdidos en todo el dataset.

Conclusiones

En la primera parte de la tarea se observó que los tres libros abordan la ciencia de aprender y reconocer patrones en los datos, pero desde perspectivas ligeramente diferentes. "Data Mining" es más práctico, buscando herramientas y técnicas para extraer conocimientos útiles de grandes conjuntos de datos. "Introduction to Machine Learning" se encuentra en el medio, tratando de entender y desarrollar algoritmos que pueden aprender de los datos, mientras que "Pattern Recognition" es más teórico, centrándose en la identificación y clasificación precisa de patrones.

A pesar de estas diferencias, hay una superposición considerable en términos de técnicas y métodos, lo que refleja la naturaleza interconectada de estas disciplinas. Las semejanzas en temas clave muestran que, aunque cada disciplina tiene sus propios objetivos y enfoques, hay una base común de conocimientos y técnicas que son relevantes en todos estos campos.

En la segunda parte de la tarea, se identificó un dataset específico, "flags", que es inherentemente multi-label. Aunque inicialmente no es adecuado para la clasificación de una sola etiqueta, con las modificaciones adecuadas, como eliminar patrones repetidos, se puede adaptar para propósitos de single-label classification. Este ejercicio no solo destaca la importancia de comprender y preparar adecuadamente los datos antes del análisis, sino que también subraya el hecho de que el reconocimiento de patrones es una disciplina versátil y adaptable, capaz de manejar una variedad de desafíos y escenarios de datos.

Referencias bibliográficas

[1] Han, J., et al. (2012). *Data mining: Concepts and Techniques, Third Edition.* Waltham: Morgan Kaufmann.

[2] Alpaydin, E. (2020). *Introduction to Machine Learning, Fourth Edition*. MIT Press.

[3] Koutroumbas, K. & Theodoridis, S. (2008). *Pattern Recognition, Second Edition*. New York: Springer.

[4] *Multilabel datasets*. (s. f.). https://mulan.sourceforge.net/datasets-mlc.html