INSTITUTO POLITÉCNICO NACIONAL

CENTRO DE INVESTIGACIÓN EN COMPUTACIÓN

Tarea 02

Datos

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 vasto mundo del Reconocimiento de Patrones (RP) nos brinda una amplia gama de técnicas y herramientas que pueden ser aplicadas a diversos conjuntos de datos para descubrir patrones ocultos y obtener conocimientos valiosos.

Los paradigmas supervisados y no supervisados son fundamentales en RP, con cada uno ofreciendo enfoques distintivos para analizar los datos. Mientras que el paradigma supervisado se centra en la predicción y clasificación basada en datos etiquetados, el no supervisado se sumerge en la identificación de estructuras intrínsecas y agrupaciones en el conjunto de datos.

La elección entre estos paradigmas y las tareas asociadas, como la clasificación y el agrupamiento, depende en gran medida de la naturaleza de los datos y de los objetivos de análisis. En esta tarea, nos embarcaremos en un viaje de exploración a través de diversos datasets, abordando tareas específicas de RP y profundizando en las características y atributos que los componen.

Desarrollo y discusión

Propósito de la tarea

Ejemplificar los tipos de atributos que aparecen en los patrones de los datasets que se usan en las diferentes tareas del RP; para el caso de la tarea classification, expresar las diferencias entre los problemas biclase y los problemas multiclase.

Parte 1

Buscar e identificar un dataset (no estudiado en clase) para clustering.

1.a) Reportar el nombre del dataset y su cardinalidad.

Un dataset para clustering que se encuentra en KEEL es Solar Flare. El dataset describe (mediante 11 características, todas de tipo nominal) determinados tipos de erupciones solares ocurridas en un periodo de 24 horas. Cada instancia representa características capturadas para una región activa en el sol [1].

Este dataset tiene 1066 patrones (instancias).

Solar Flare data set			
Type	Unsupervised	**Origin**	Real world
Features	11	**(Real / Integer / Nominal)**	(0 / 0 / 11)
Instances	1066	**Missing values?**	No

Tabla 1: Información general de Solar Flare dataset

1.b) Reportar los atributos en el formato de KEEL.

Attribute	**Domain**	**Attribute**	**Domain**
LargestSpotSize	{A, R, S, X, K, H}	BecomeHist	{1, 2}
SpotDistribution	{X, O, I, C}	Area	{1, 2}
Activity	{1, 2}	C-class	{0, 1, 2, 3, 4, 5, 6, 7, 8}
Evolution	{1, 2, 3}	M-class	{0, 1, 2, 3, 4, 5}
Prev24Hour	{1, 2, 3}	X-class	{0, 1, 2}
HistComplex	{1, 2}		

Tabla 2: Descripción de atributos de Solar Flare dataset

1.c) Reportar el ejemplo de un patrón del dataset y explicar el significado del valor que aparece en el primer atributo.

Revisemos el primer patrón: A, X, 1, 3, 1, 1, 1, 1, 0, 0, 0, H. El primer atributo representa que la mancha solar más grande es de categoría A.

1.d) Redactar un párrafo donde se expresen las razones por las que este dataset corresponde al paradigma no supervisado (pista: ¿los patrones del dataset se asocian con algo?)

El dataset tiene 11 atributos de entrada, pero no un atributo de salida con el que se asocie cada patrón. Esto implica que, en lugar de emplearlo para propósitos predictivos, su orientación es hacia lo descriptivo. Por lo tanto, este dataset corresponde al paradigma no supervisado.

1.e) Explicar por qué este dataset se puede usar en la tarea Recall.

El conjunto de datos "Solar Flare" es adecuado para la tarea "recall" porque es posible asociar patrones del dataset con otros patrones. Esto nos permite recuperar eventos con características parecidas, entender las tendencias históricas y hacer inferencias sobre eventos futuros.

Parte 2

Buscar e identificar un dataset (no estudiado en clase) para regression.

2.a) Reportar el nombre del dataset y su cardinalidad.

Un dataset para regression que se encuentra en KEEL es Forest Fires. Se trata de una tarea de regresión, en la que el objetivo es predecir el área quemada de los incendios forestales, en la región noreste de Portugal, mediante el uso de datos meteorológicos y de otro tipo [2].

El dataset tiene 517 patrones (instancias).

Type	**Regression**	**Origin**	**Real world**
Features	**12**	**(Real / Integer / Nominal)**	**(7 / 5 / 0)**
Instances	**517**	**Missing values?**	**No**

Tabla 3: Información general de Forest Fires dataset.

2.b) Reportar los atributos en el formato de KEEL.

Attribute	**Domain**	**Attribute**	**Domain**
X	[1,9]	DC	[7.9,860.6]
Y	[2,9]	ISI	[0,56.1]
Month	[1,12]	Temp	[2.2,33.3]
Day	[1,7]	RH	[15,100]
FFMC	[18.7,96.2]	Wind	[0.4,9.4]
DMC	[1.1,291.3]	Rain	[0,6.4]
Area	[0,1090.84]		

Tabla 4: Descripción de atributos de Forest Fires dataset.

2.c) Reportar el ejemplo de un patrón del dataset y explicar el significado del valor que aparece en el segundo atributo.

El segundo atributo Y hace referencia a una coordenada espacial vertical en un sistema de grilla.

7, 4, 3, 1, 90.1, 39.7, 86.6, 6.2, 16.1, 29, 3.1, 0, 1.75

2.d) ¿En qué posición aparece el valor real que se asocia con el patrón?

Se encuentra en la posición 13 y se refiere al área de exploración.

2.e) Redactar un párrafo donde se expresen las razones por las que este dataset corresponde al paradigma supervisado (pista: ¿los patrones del dataset se asocian con algo?)

Este dataset corresponde al paradigma supervisado porque se asocian valores reales a los patrones. Hay que notar que contiene únicamente atributos numéricos (tipo Real o Integer). Esto permite realizar regresión sobre el conjunto de datos.

Parte 3

Buscar e identificar un dataset (no estudiado en clase) para classification biclase.

3.a) Reportar el nombre del dataset.

Un dataset para classification biclase que se encuentra en KEEL es Mushroom. Este conjunto de datos incluye descripciones de muestras hipotéticas correspondientes a 23 especies de setas branquiales de las familias Agaricus y Lepiota. Cada especie se identifica como comestible o venenosa [3].

Type	**Classification**	**Origin**	**Real world**
Features	22	(Real / Integer / Nominal)	(0 / 0 / 22)
Instances	8124	Classes	2
Missing values?			Yes

Tabla 5: Información general de Mushroom dataset.

3.b) Reportar la cardinalidad del dataset y las cardinalidades de las clases.

El dataset tiene 2001 instancias y 22 atributos de tipo real. Ya que es biclase, la cardinalidad de las clases es dos.

3.c) Reportar los atributos en el formato de KEEL.

Attribute	**Domain**	**Attribute**	**Domain**	**Attribute**	**Domain**
Cap-shape	{x, b, s, f, k, c}	Gill-color	{k, n, g, p, w, h, u, e, b, r, y, o}	Veil-type	{p, u}
Cap-surface	{s, y, f, g}	Stalk-shape	{e, t}	Veil-color	{w, n, o, y}
Cap-color	{n, y, w, g, e, p, b, u, c, r}	Stalk-root	{e, c, b, r, u, z}	Ring-number	{o, t, n}

Bruises	{t, f}	Stalk-surface-above-ring	{s, f, k, y}	Ring-type	{p, e, l, f, n, c, s, z}
Odor	{p, a, l, n, f, c, y, s, m}	Stalk-surface-below-ring	{s, f, y, k}	Spore-print-color	{k, n, u, h, w, r, o, y, b}
Gill-attachment	{f, a, d, n}	Stalk-color-above-ring	{w, g, p, n, b, e, o, c, y}	Population	{s, n, a, v, y, c}
Gill-spacing	{c, w, d}	Stalk-color-below-ring	{w, p, g, b, n, e, y, o, c}	Habitat	{u, g, m, d, p, w, l}
Gill-size	{n, b}	Class	{p, e}		

Tabla 6: Descripción de atributos de Mushroom dataset.

3.d) ¿En qué posición aparece el atributo de clase?

Aparece en la posición 23, hace referencia a si el hongo es comestible o venenoso.

3.e) Reportar el ejemplo de un patrón del dataset y explicar el significado del valor que aparece en el tercer atributo.

x,s,n,t,p,f,c,n,k,e,e,s,s,w,w,p,w,o,p,k,s,u,p

El tercer atributo se relaciona al color de la parte superior del hongo, en este caso “navy”, es decir azul marino.

3.f) ¿Cuál es la etiqueta de clase del patrón de ejemplo del inciso previo?

Fue “p”, es decir, se trata de un hongo venenoso.

3.g) Redactar un párrafo donde se expresen las razones por las que este dataset corresponde al paradigma supervisado (pista: ¿los patrones del dataset se asocian con algo?)

En este dataset se asigna a cada patrón del dataset una categoría del atributo de clase, este valor es conocido como etiqueta de clase. De esta manera, si queremos clasificar cada patrón en alguna de las dos clases que se tienen, es posible conocer en qué forma los atributos las caracterizan.

Parte 4

Buscar e identificar un dataset (no estudiado en clase) para classification multiclase

4.a) Reportar el nombre del dataset.

Un dataset para classification multiclase de KEEL es Contraceptive Method Choice. Tiene su origen en la Encuesta nacional de prevalencia del uso de anticonceptivos en Indonesia de 1987. Las muestras son mujeres casadas que no estaban embarazadas o no sabían si lo estaban en el momento de la entrevista. El problema consiste en predecir la elección actual de método anticonceptivo (no usa, métodos de larga duración o métodos de corta duración) de una mujer en función de sus características demográficas y socioeconómicas [4].

4.b) Reportar la cardinalidad del dataset y la cardinalidad de las clases.

El dataset tiene 1473 instancias y 9 atributos de tipo entero. La cardinalidad de su atributo de clase es 3.

Contraceptive Method Choice data set			
Type	**Classification**	**Origin**	**Real world**
Features	9	(Real / Integer / Nominal)	(0 / 9 / 0)
Instances	1473	Classes	3
Missing values?	No		

Tabla 7: Información general de Contraceptive Method Choice dataset.

4.c) Reportar los atributos en el formato de KEEL.

Attribute	Domain
Wife_age	[16,49]
Wife_education	[1,4]
Husband_education	[1,4]
Children	[0,16]
Wife_religion	[0,1]
Wife_working	[0,1]
Husband_occupation	[1,4]
Standard-of-living	[1,4]
Media_exposure	[0,1]
Contraceptive_method	{1,2,3}

Tabla 8: Descripción de atributos de Contraceptive Method Choice dataset.

4.d) ¿En qué posición aparece el atributo de clase?

Aparece en la posición 10, hace referencia a si la mujer no usa método anticonceptivo, o si es de corto plazo o largo plazo.

4.e) Reportar el ejemplo de un patrón del dataset y explicar el significado del valor que aparece en el segundo atributo.

El último patrón del dataset es 17,3,3,1,1,1,2,4,0,3

El segundo atributo se refiere al nivel educativo de la mujer que en este caso es 3, es decir, secundaria.

4.f) ¿Cuál es la etiqueta de clase del patrón del ejemplo del inciso previo?

Es 3, esto significa que la mujer utiliza un método anticonceptivo de largo plazo.

Parte 5

Buscar e identificar un dataset para classification biclase con patrones binarios (un patrón binario es el que contiene únicamente atributos binarios; un atributo binario es el que toma el valor 1, o bien, el valor 0).

5.a) Reportar el ejemplo de un patrón del dataset y explicar el significado de los valores de todos los atributos.

El dataset es SPECT Heart de UCI [5].

Atributo	Valor
OVERALL_DIAGNOSIS	**1**
F1	0
F2	0
F3	0
F4	1
F5	0
F6	0
F7	0
F8	1
F9	1
F10	0
F11	0
F12	0
F13	1
F14	1
F15	0
F16	0
F17	0
F18	0
F19	0
F20	0
F21	0
F22	0

Tabla 9: Ejemplo de un patrón del conjunto de datos.

Significado de los valores de los atributos:

OVERALL_DIAGNOSIS: Es la columna de clasificación que indica el diagnóstico general. Un valor de "1" indica un diagnóstico anormal y un valor de "0" indica un diagnóstico normal.

F1 a F22: Estos atributos representan diagnósticos parciales derivados de imágenes cardíacas SPECT. Cada atributo es binario, donde un valor de "1" indica la presencia de una característica específica y un valor de "0" su ausencia. Estas características parciales son

reducciones binarias de características originales continuas obtenidas de las imágenes SPECT.

El conjunto de datos "SPECT" describe el diagnóstico de imágenes cardíacas usando la técnica Single Proton Emission Computed Tomography (SPECT). El propósito es clasificar a los pacientes en dos categorías: normal y anormal, utilizando 22 características binarias derivadas de las imágenes. En el patrón mostrado anteriormente, el diagnóstico general (OVERALL_DIAGNOSIS) es "anormal" basado en las características binarias presentes.

Conclusiones

Después de una exploración detallada de varios datasets y de adentrarnos en las tareas específicas del Reconocimiento de Patrones, hemos adquirido una comprensión más profunda de las sutilezas y matices que cada dataset presenta.

Se han explorado desde datasets diseñados para diferentes tareas hasta las diferencias entre clasificación biclase y multiclase. También se ha ampliado nuestro horizonte de conocimientos en el dominio del RP. La exploración de repositorios como KEEL ha sido fundamental, proporcionando mayor entendimiento de la complejidad de los datos.

Con esto en mente, queda claro que el análisis detallado y la interpretación adecuada de los conjuntos de datos son esenciales para cualquier tarea de RP exitosa.

A través de este ejercicio, no sólo hemos ganado habilidades analíticas, sino también una apreciación más profunda de la importancia de una exploración cuidadosa y metódica en el campo del Reconocimiento de Patrones.

Referencias bibliográficas

[1] KEEL: a software tool to assess evolutionary algorithms for data mining problems (regression, classification, clustering, pattern mining and so on). (s. f.). https://sci2s.ugr.es/keel/dataset.php?cod=1295

[2] KEEL: a software tool to assess evolutionary algorithms for data mining problems (regression, classification, clustering, pattern mining and so on). (s. f.-b). https://sci2s.ugr.es/keel/dataset.php?cod=19

[3] KEEL: a software tool to assess evolutionary algorithms for data mining problems (regression, classification, clustering, pattern mining and so on). (s. f.-c). https://sci2s.ugr.es/keel/dataset.php?cod=178

[4] KEEL: a software tool to assess evolutionary algorithms for data mining problems (regression, classification, clustering, pattern mining and so on). (s. f.-d). https://sci2s.ugr.es/keel/dataset.php?cod=58

[5] UCI Machine Learning Repository. (s. f.-b). https://archive.ics.uci.edu/dataset/95/spect+heart