Práctica 1:

Análisis Predictivo Empresarial Mediante Clasificación

Curso 2019/2020

SOFÍA ALMEIDA BRUNO
sofialmeida@correo.ugr.es

Grupo IN 2

Jueves 9:30-10:30

Índice

1. Introducción

En este práctica se abordará un problema de clasificación del mundo real para, mediante el uso de los algoritmos de clasificación supervisada vistos en clase de teoría y las herramientas y recursos expuestos en clase de prácticas, realizar una predicción sobre el mismo y analizar cómo de buena es esta clasificación. Se compararán distintos algoritmos y se examinará la predicción obtenida en función a los mismos según distintos criterios de precisión.

El problema con el que se trabajará proviene de la plataforma "Driven data", usa los datos de "Taarifa" (API web libre que está trabajando en un poryecto de innovación en Tanzania) y del Ministerio de Agua de Tanzania. El objetivo es predecir qué bombas de agua funcionan, cuáles necesitan alguna reparación y cuáles están rotas. Es decir, estamos ante un problema de clasificación con tres clases diferentes. Se trata de predecir mediante variables como: qué tipo de bomba es, cuándo se instaló, cantidad de agua disponible,... ante qué tipo de bomba de agua nos encontramos. Saber qué puntos de agua fallarán permitirá mejorar las tareas de mantenimiento y asegurar que hay agua limpia y potable disponible para las comunidades de Tanzania.

Abordaremos el problema a partir de un conjunto de datos formado por 59.400 instancias, de las cuales conocemos información sobre 39 variables, además de su clase (una de las tres ya mencionadas). Usando el nodo `Data Explorer` podemos hacer una exploración inicial del conjunto. En primer lugar, observamos la frecuencia de las clases: de todas las instancias 32259 son bombas de agua funcionales, 22824 no funcionales y 4317 funcionales pero necesitan una reparación. Las clases están desbalanceadas, observamos una gran diferencia en el número de ejemplos de bombas funcionales y aquellas que pese a ser funcionales requieren mantenimiento.

Nada más cargar el fichero observamos que es un conjunto de datos que posee valores perdidos, además de algunos valores "unknown".

Toda la experimentación se realizará usando una validación cruzada de 5 particiones. La semilla aleatoria empleada en aquellos algoritmos que lo requieran será: 12345. Los experimentos realizados en esta práctica se ejecutaron en un ordenador con sistema operativo Ubuntu 16.04 con procesador Intel(R) Core(TM) i5-2410M CPU @ 2.30GHz.

Se utilizará la validación cruzada en la ejecución de todos los algoritmos, mediante los nodos `X-Partitioner` y `X-Aggregator` de KNIME. Se configuran para crear 5 particiones, luego en cada experimento se utilizará un conjunto de entrenamiento de tamaño 47520 y un conjunto de prueba formado por 11.880 instancias.

Hemos visto en clase que comparar los algoritmos solo por la precisión que consiguen en la predicción no es suficiente, ya que en conjuntos desbalanceados malos algoritmos podrían obtener una alta precisión. Así, utilizaremos medidas sensibles al desbalanceo. Se siguió el tutorial proporcionado por el profesor de

prácticas sobre cómo comparar diferentes algoritmos para obtener las tablas de resultados.

2. Resultados obtenidos

3. ZeroR

Para comenzar (y sin incluirlo como algoritmo a estudiar), he decidido observar el comportamiento del clasificador ZeroR. Este clasificador predice que cualquier instancia pertenecerá a la clase mayoritaria. Aunque ya sabemos que no obtendremos un buen resultado utilizando este clasificador porque solo clasificará correctamente las instancias que verdaderamente pertenezcan a la clase mayoritaria, nos servirá para tener una cota inferior de las medidas. Si en algún momento durante el desarrollo de la práctica obtuvieramos resultados peores que los obtenidos con este clasificador sospecharemos que estamos haciendo algo mal.

Podemos observar el metanodo creado en KNIME para este algoritmo en la Figura 1. Se ha utilizado el nodo `ZeroR` de `Weka`.

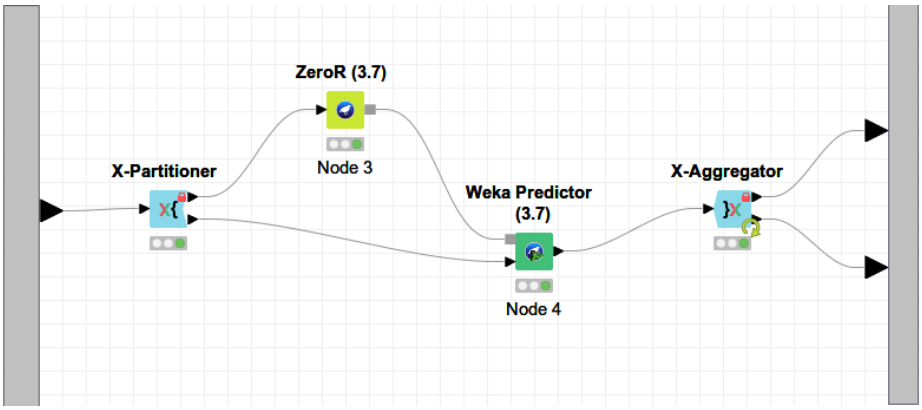

Figura 1: Metanodo ZeroR

En la Tabla 3 se encuentran las medidas de precisión obtenidas con este algoritmo. En este caso, conociendo la distribución de las clases, se podrían haber calculado manualmente sin necesidad de ejecutar el algoritmo, pues por su naturaleza el número de verdaderos positivos (TP) se corresponderá al número

Tabla 1: ZeroR - criterios de precisión

	TP	FP	TN	FN	TPR	TNR	PPV	Accur.	F1-score	G-mean
ZeroR	32259	27414	0	0	1	0	0.543	0.543	0.704	0

3.1. C4.5

3.2. Resultados obtenidos

3.2.1. 1-NN

Tabla 2: Colposcopy

	Tasa clasificación	Tasa reducción	Agregado	Tiempo (μs)
Partición 1	74.5763	0	37.2881	4316
Partición 2	70.1754	0	35.0877	3613
Partición 3	73.6842	0	36.8421	5636
Partición 4	75.4386	0	37.7193	3638
Partición 5	82.4561	0	41.2281	3411
Media	75.2661	0	37.6331	4122.8

Tabla 3: Ionosphere

	Tasa clasificación	Tasa reducción	Agregado	Tiempo (μs)
Partición 1	90.1408	0	45.0704	3804
Partición 2	80	0	40	3470
Partición 3	82.8571	0	41.4286	3319
Partición 4	92.8571	0	46.4286	3938
Partición 5	87.1429	0	43.5714	3443
Media	86.5996	0	43.2998	3594.8

Tabla 4: Texture

	Tasa clasificación	Tasa reducción	Agregado	Tiempo (μs)
Partición 1	93.6364	0	46.8182	9351
Partición 2	89.0909	0	44.5455	8895
Partición 3	94.5455	0	47.2727	8251
Partición 4	92.7273	0	46.3636	7658
Partición 5	92.7273	0	46.3636	7524
Media	92.5455	0	46.2727	8335.8

3.2.2. Resultados globales

Tabla 5: Colposcopy

Algoritmo	Tasa clasificación	Tasa reducción	Agregado	Tiempo (μs)
1-NN	75.2661	0	37.6331	4122.8
RELIEF	75.9798	36.4516	56.2157	12357.8
BL	76.2831	14.5161	45.3996	7.07349e+06

Tabla 6: Ionosphere

Algoritmo	Tasa clasificación	Tasa reducción	Agregado	Tiempo (μs)
1-NN	86.5996	0	43.2998	3594.8
RELIEF	87.4567	2.94118	45.199	10962.6
BL	86.0402	20.5882	53.3142	5.62155e+06

Tabla 7: Texture

Algoritmo	Tasa clasificación	Tasa reducción	Agregado	Tiempo (μs)
1-NN	92.5455	0	46.2727	8335.8
RELIEF	93.0909	5.5	49.2955	19242.4
BL	90.7273	22.5	56.6136	1.40398e+07

4. Análisis de resultados

5. Configuración de algoritmos

6. Procesado de datos

7. Interpretación de resultados

8. Contenido adicional

9. Bibliografía