Actividad 1. Comparando métodos de detección de anomalías

UCO ONLINE

0. Datos personales

Indica tu nombre y apellidos en el cuadernillo

Nombre y Apellidos: Juan José Méndez Torrero

1. Seleccionar conjunto de datos (Fase 1)

Dentro de los conjuntos disponibles en la web: http://odds.cs.stonybrook.edu/

Selecciona tres conjuntos de datos que pertenezcan a conjuntos de datos de puntos multivariantes.

Rellena en la tabla que se muestra la información de los tres conjuntos de datos que se han seleccionado para el estudio comparativo. En esta tabla se indica: el nombre del conjunto de datos (#dataset), el número de ejemplos/instancias que tienen (#ejemplos), cuántos atributos tiene para representar cada instancia (#dimensiones) y el porcentaje de anomalías que tienen (#anomalías).

#Dataset	**#Ejemplos**	**#Dimensiones**	**#Anomalías(%)**
Pima	768	8	268 (35%)
Satellite	6435	36	2036 (32%)
Shuttle	49097	9	3511 (7%)

Recuerde que en el estudio tienes que particionar cada conjunto de datos en training y test utilizando holdout con una proporción 70% para entrenamiento y 30% para test y los normalizarás (lo tienes considerado en los ficheros de ejemplo que tienes que usar).

2. Configurar parámetros de los algoritmos (Fase 2)

Selecciona dos algoritmos de los estudiados para configurar sus parámetros. Basándote en los ficheros disponibles:

- ***Configurar_algoritmo_KNN.ipynb***: tiene el código disponible para llevar a cabo la configuración del algoritmo KNN de acuerdo con el número de vecinos.
- ***Configurar_algoritmo_LOF.ipynb***: tiene el código disponible para llevar a cabo la configuración del algoritmo LOF de acuerdo con el número de vecinos.
- ***Configurar_algoritmo_COF.ipynb***: tiene el código disponible para llevar a cabo la configuración del algoritmo COF de acuerdo con el número de vecinos.
- ***Configurar_algoritmo_CBLOF.ipynb***: tiene el código disponible para llevar a cabo la configuración del algoritmo CBLOF de acuerdo con el número de clústeres.
- ***Configurar_algoritmo_HBOS.ipynb***: tiene el código disponible para llevar a cabo la configuración del algoritmo HBOS de acuerdo con el número de barras.

Incluye las tablas que proporciona los notebooks Configurar_algoritmo_X y Configurar_algoritmo_Y (siendo X e Y los dos algoritmos seleccionados) para justificar el valor del parámetro que se estudia en cada uno de ellos, para cada uno de los conjuntos de datos.

Algoritmo	Datos	#Ejemplos	#Dimensiones	Anomalías(%)	*k*	Se	Sp
kNN	pima	768	8	34.8958	5	0.4819	0.7432
kNN	pima	768	8	34.8958	6	0.4699	0.7432
kNN	pima	768	8	34.8958	7	0.4819	0.7365
kNN	pima	768	8	34.8958	8	0.494	0.75
kNN	**pima**	**768**	**8**	**34.8958**	**9**	**0.506**	**0.75**
kNN	pima	768	8	34.8958	10	0.506	0.7432
kNN	pima	768	8	34.8958	11	0.506	0.7432
kNN	pima	768	8	34.8958	12	0.494	0.75
kNN	pima	768	8	34.8958	13	0.494	0.75

Algoritmo	Datos	#Ejemplos	#Dimensiones	Anomalías(%)	*N-cluster*	Se	Sp
CBLOF	pima	768	8	34.8958	15	0.5301	0.6419
CBLOF	pima	768	8	34.8958	16	0.5663	0.6757
CBLOF	pima	768	8	34.8958	17	0.5181	0.6284
CBLOF	pima	768	8	34.8958	18	0.5904	0.6419
CBLOF	pima	768	8	34.8958	19	0.5542	0.6216
CBLOF	pima	768	8	34.8958	20	0.5542	0.6486
CBLOF	pima	768	8	34.8958	21	0.6024	0.6622
CBLOF	**pima**	**768**	**8**	**34.8958**	**22**	**0.6145**	**0.6554**

Algoritmo	Datos	#Ejemplos	#Dimensiones	Anomalías(%)	k	Se	Sp
kNN	satellite	6435	36	31.6395	200	0.6382	0.8343
kNN	**satellite**	**6435**	**36**	**31.6395**	**210**	**0.6432**	**0.8358**
kNN	satellite	6435	36	31.6395	220	0.6365	0.8351
kNN	satellite	6435	36	31.6395	230	0.6365	0.8313
kNN	satellite	6435	36	31.6395	240	0.6348	0.8321
kNN	satellite	6435	36	31.6395	250	0.6348	0.8328
kNN	satellite	6435	36	31.6395	260	0.6332	0.8343
kNN	satellite	6435	36	31.6395	270	0.6265	0.8358
kNN	satellite	6435	36	31.6395	280	0.6265	0.8336

Algoritmo	Datos	#Ejemplos	#Dimensiones	Anomalías(%)	N-cluster	Se	Sp
CBLOF	satellite	6435	36	31.6395	5	0.5611	0.7961
CBLOF	satellite	6435	36	31.6395	6	0.6248	0.8223
CBLOF	satellite	6435	36	31.6395	7	0.5712	0.7976
CBLOF	satellite	6435	36	31.6395	8	0.5544	0.7969
CBLOF	**satellite**	**6435**	**36**	**31.6395**	**9**	**0.6298**	**0.8321**
CBLOF	satellite	6435	36	31.6395	10	0.5879	0.8066
CBLOF	satellite	6435	36	31.6395	11	0.5226	0.7646
CBLOF	satellite	6435	36	31.6395	12	0.5343	0.7684

Algoritmo	Datos	#Ejemplos	#Dimensiones	Anomalías(%)	k	Se	Sp
kNN	shuttle	49097	9	7.1511	5	0.2015	0.9442
kNN	shuttle	49097	9	7.1511	6	0.2072	0.9425
kNN	shuttle	49097	9	7.1511	7	0.1958	0.9458
kNN	shuttle	49097	9	7.1511	8	0.2025	0.9446
kNN	shuttle	49097	9	7.1511	9	0.2072	0.9432
kNN	**shuttle**	**49097**	**9**	**7.1511**	**10**	**0.2082**	**0.9438**
kNN	shuttle	49097	9	7.1511	11	0.2072	0.9436
kNN	shuttle	49097	9	7.1511	12	0.2025	0.9423

Algoritmo	Datos	#Ejemplos	#Dimensiones	Anomalías(%)	N-cluster	Se	Sp
CBLOF	shuttle	49097	9	7.1511	10	0.2956	0.9477
CBLOF	shuttle	49097	9	7.1511	11	0.3203	0.9481
CBLOF	shuttle	49097	9	7.1511	12	0.2937	0.9471
CBLOF	shuttle	49097	9	7.1511	13	0.3964	0.9545
CBLOF	**shuttle**	**49097**	**9**	**7.1511**	**14**	**0.4838**	**0.9622**
CBLOF	shuttle	49097	9	7.1511	15	0.0998	0.9343
CBLOF	shuttle	49097	9	7.1511	16	0.0998	0.9343
CBLOF	shuttle	49097	9	7.1511	17	0.0989	0.9343

Tras la justificación, mostrando tablas similares a la indicadas en la última lección vista esta semana y donde se analizan varios valores del parámetro estudiado. Incluye la información final en la siguiente tabla:

Dataset	Algoritmo	Parámetro/Valor
Pima	KNN	9
Pima	CBLOF	22
Satellite	KNN	210
Satellite	CBLOF	9
Shuttle	KNN	10
Shuttle	CBLOF	14

3. Ejecutar los algoritmos con el conjunto de datos (Fase 3)

Basándote en el notebook: "Configurar_modelos_anomalias.ipynb" debes modificarlo para considerar tus conjunto de datos, tus dos algoritmos y los parámetros obtenidos del paso anterior. Muestra la tabla de resultados (como la que se indica) para tiempo de cómputo, sensibilidad, especificidad, precisión y auc-roc.

Tiempo de cómputo

Datos	#Ejemplos	#Dimensiones	Anomalías(%)	CBLOF	KNN
pima	768	8	34.8958	2.2637	0.0316
satellite	6435	36	31.6395	0.3456	1.8368
shuttle	49097	9	7.1511	1.0051	8.7486

Sensibilidad

Datos	#Ejemplos	#Dimensiones	Anomalías(%)	CBLOF	KNN
pima	768	8	34.8958	0.6024	0.506
satellite	6435	36	31.6395	0.5917	0.614
shuttle	49097	9	7.1511	0.3954	0.2073

Especificidad

Datos	#Ejemplos	#Dimensiones	Anomalías(%)	CBLOF	KNN
pima	768	8	34.8958	0.6216	0.75
satellite	6435	36	31.6395	0.8413	0.8474
shuttle	49097	9	7.1511	0.954	0.9378

Precisión

Datos	#Ejemplos	#Dimensiones	Anomalías(%)	CBLOF	KNN
pima	768	8	34.8958	0.4717	0.5316
satellite	6435	36	31.6395	0.6419	0.6592
shuttle	49097	9	7.1511	0.3969	0.2032

AUC-ROC

Datos	#Ejemplos	#Dimensiones	Anomalías(%)	CBLOF	KNN
pima	768	8	34.8958	0.612	0.628
satellite	6435	36	31.6395	0.7165	0.7307
shuttle	49097	9	7.1511	0.6747	0.5725

4. Análisis de los resultados (Fase 4)

Lleva a cabo un análisis de los resultados obtenidos e indica que método utilizarías para resolver los problemas en función de los resultados obtenidos. Si es necesario, puedes indicar el mejor algoritmo según el conjunto de datos.

Tras haber ejecutado los algoritmos CBLOF y KNN sobre los tres conjuntos de datos seleccionados, se puede observar que, en tiempo de cómputo, cuando el conjunto de datos cuenta con un gran número de instancias, el algoritmo tiende a tardar más en encontrar el agrupamiento correcto.

Con respecto a la sensibilidad obtenida en ambos algoritmos, se puede observar que CBLOF consigue determinar, con mayor porcentaje de acierto, las instancias anómalas para dos de los conjuntos de datos seleccionados. En el caso en el que el algoritmo KNN ha ofrecido mejores resultados, la diferencia entre la sensibilidad obtenida para ambos algoritmos es pequeña. Por consecuente, según estos resultados, nos decantaríamos por el algoritmo CBLOF, ya que ofrece, en general, mejores resultados a la hora de detectar anomalías.

En el caso de la especificidad, observamos que KNN ofrece mejores resultados para dos de los conjuntos de datos seleccionados, siendo en el conjunto de datos *Pima* en el que el algoritmo KNN consigue clasificar, en mayor porcentaje, las instancias normales del conjunto de datos. En este caso, el algoritmo CBLOF consigue buenos resultados, aunque es en el conjunto de datos *Pima* donde no consigue clasificar correctamente las instancias normales.

La precisión obtenida en ambos algoritmos es bastante similar para los tres conjuntos de datos seleccionados, siendo el algoritmo KNN el que consigue una mayor precisión para dos de los conjuntos de datos, aunque en el caso en el que el conjunto de datos tiene un mayor número de instancias, el algoritmo KNN funciona bastante peor que el algoritmo CBLOF.

Finalmente, se puede observar que el algoritmo KNN clasifica con mayor probabilidad dos de conjuntos de datos seleccionados. Aunque, para el conjunto de datos con mayor número de instancias, la probabilidad de clasificar correctamente las instancias es mucho mayor usando el algoritmo CBLOF.