Universidad de Valladolid

Ingeniería Informática

Minería de datos

Evaluación de varios métodos

Autor:
Vicente Martínez Franco

28 de noviembre de 2016

Índice

1. Introducción

En esta práctica se busca comparar el comportamiento de varios métodos de aprendizaje sobre varios conjuntos de entrenamiento para evaluar si alguno de ellos tiene un mejor comportamiento.

Los métodos elegidos para ello son los siguientes:

- SVM con kernel lineal (SMO)
- 3-NN (Ibk)
- Naive Bayes
- J48

Y se evaluaran sobre los siguientes 10 conjuntos de datos, los 7 primeros pertenecientes a los datos disponibles en Weka y los 3 restantes a datos extraídos del UCI:

1. SoyBean
2. Ionosphere
3. Vote
4. Diabetes
5. Labor
6. Glass
7. Segment-test
8. Arrythmia
9. Dbworld, en concreto el fichero dbworld-subjects
10. Post-operative

Lo primero que realizaremos será un test de signos sobre los clasificadores SVM y J48. Después procederemos a la evaluación mediante rankings utilizando ya para esto los 4 métodos escogidos.

2. Test de signos

El test de signos se basa en comparar el número de victorias que obtiene cada algoritmo en los conjuntos de datos. En el caso de que ambos algoritmos sean equivalentes ambos obtendrán aproximadamente la mitad de las victorias.

Para la realización de este test no deben contarse únicamente la victorias significativas sino que han de tenerse todas en cuenta, ya que, aunque la victoria en un conjunto de datos pueda ser aleatoria, la probabilidad de que esto suceda para multiples conjuntos de datos se reduce significativamente.

2.1. Realización del test

Aquí podemos ver los datos obtenidos durante la realización del test. En primer lugar observamos la tabla con el porcentaje de aciertos y fallos obtenido en cada algoritmo:

Datos	SMO		J48		Ganador
	Correctas	Incorrectas	Correctas	Incorrectas	
SoyBean	93.8507 %	6.1493 %	91.5081 %	8.4919 %	SOM
Ionosphere	88.604 %	11.396 %	91.453 %	8.547 %	J48
Vote	96.092 %	3.908 %	96.3218 %	3.6782 %	J48
Diabetes	77.3438 %	22.6563 %	73.8281 %	26.1719 %	SOM
Labor	89.4737 %	10.5263 %	73.6842 %	26.3158 %	SOM
Glass	56.0748 %	43.9252 %	66.8224 %	33.1776 %	J48
Segment-test	92.2222 %	7.7778 %	93.4568 %	6.5432 %	J48
Arrythmia	70.1327 %	29.8673 %	64.823 %	35.177 %	SOM
Dbworld	85.9375 %	14.0625 %	71.875 %	28.125 %	SOM
Post-operative	68.8889 %	31.1111 %	70 %	30 %	J48

Lo siguiente que haremos es contar el número de victorias de cada uno de los algoritmos.

	SOM	J48
Número de victorias	5	5

2.2. Resultados

En este caso se ha producido un empate en el número de victorias de ambos algoritmos con lo cual podemos aceptar la hipótesis nula que nos dice que ambos algoritmos tienen un comportamiento similar para el conjunto de datos dado.

Esto sin embargo es sobre todo debido a la gran diversidad de los datos escogidos para realizar el test, que no tiene nada en común entre si.

3. Rankings

En el método de evaluación basado en rankings lo que hacemos es ordenar los algoritmos de mejor a peor para cada uno de los conjuntos de datos evaluados. En el caso de que se produzca un empate se les da un valor promedio a los algoritmos que hayan empatado.

Después promediamos el resultado para cada método y los ordenamos según este promedio. A partir de aquí pueden aplicarse varios métodos estadísticos. En nuestro caso realizaremos el test de Iman y Davenport y en caso de encontrar diferencias significativas aplicaremos un test post-hoc para determinar cuales de ellos son diferentes.

3.1. Realización de los rankings

En primer lugar vemos la tabla con el ranking según la tasa de error obtenida

Datos	SVM (SOM)	J48	Naive Bayes	3-NN (Ibk)
SoyBean	6.1493 % (1)	8.4919 % (3)	7.0278 % (2)	8.7848 % (4)
Ionosphere	11.396 % (2)	8.547 % (1)	17.3789 % (4)	13.6752 % (3)
Vote	3.908 % (2)	3.6782 % (1)	9.8851 % (4)	7.5862 % (3)
Diabetes	22.6563 % (1)	26.1719 % (3)	23.6979 % (2)	29.8177 % (4)
Labor	10.5263 % (1.5)	26.3158 % (4)	10.5263 % (1.5)	17.5439 % (3)
Glass	43.9252 % (3)	33.1776 % (2)	51.4019 % (4)	29.4393 % (1)
Segment-test	7.7778 % (3)	6.5432 % (2)	13.5802 % (4)	5.3086 % (1)
Arrythmia	29.8673 % (1)	35.177 % (2)	38.2743 % (3)	47.1239 % (4)
Dbworld	14.0625 % (2)	28.125 % (4)	10.9375 % (1)	20.3125 % (3)
Post-operative	31.1111 % (2)	30 % (1)	32.2222 % (3)	46.6667 % (4)

Lo siguiente que haremos es calcular el valor promedio que ha obtenido cada uno de los algoritmos en los rankings.

	SOM	J48	Naive Bayes	Ibk
Ranking medio	1.85	2.3	2.85	3

3.2. Test de Iman y Davenport

El test de Iman y Davenport es utilizado para comprobar si existen diferencias entre los distintos métodos dado su ranking medio. Se basa en la comparación con la distribución F mediante el siguiente estadístico:

$F_F = \frac{(N-1)\chi_F^2}{N(k-1)-\chi_F^2}$

Donde:

$\chi_F^2 = \frac{12N}{k(k+1)}[\sum_j R_j^2 - \frac{k(k+1)^2}{4}]$

En este caso :

$\chi_F^2 = \frac{120}{4(5)}[\sum_j R_j^2 - \frac{4(5)^2}{4}] = 5{,}01$

$F_F = \frac{(9)5{,}10}{10(3)-5{,}01} = 1{,}804$

El valor crítico para $\alpha = 0.05$ k-1 = 3 y 27 grados de libertad es de: 2.960

Dado que el valor obtenido es menor que el valor crítico no se rechaza la hipótesis nula, con lo que los algoritmos no son significativamente distintos.

3.3. Test post-hoc

Dado que los algoritmos no son significativamente distintos no podemos realizar este test, y que no tendría sentido.

3.4. Resultados

Una vez más nos encontramos con que no existe diferencia entre los algoritmos utilizados debido probablemente a la misma razón que la primera vez, la diversidad de los conjuntos de datos utilizados.

4. Conclusiones

En esta práctica hemos podido observar el funcionamiento de algunos de los métodos de comparación de algoritmos de clasificación para varios conjuntos de datos.

En este caso debido a la gran diferencia existente entre los distintos conjuntos de datos no se han podido observar diferencias y no hemos podido llegar a realizar todos los test pero si que hemos podido ver el funcionamiento básico de la comparativa de clasificadores. Para poder obtener mejores resultados deberían utilizarse conjuntos de datos correspondientes al mismo problema.