Evaluación de Hipótesis

García Prado, Sergio
sergio@garciparedes.me

20 de marzo de 2017

Resumen

*En este documento se realizarán experimentos sobre 3 conjuntos de datos utilizados para entrenar y verificar la tasa de error obtenida mediante distintas metodologías. Los algoritmos utilizados se basan en aprendizaje supervisado para la generación de árboles de decisión (*J48*) y conjuntos de reglas(*JRIP*) aplicado a tareas de clasificación.*

1. Introducción

El motivo principal por el cual se realiza este conjunto de experimentos es la comparación de las distintas tasas de error mediante cada una de las técnicas, tratando de apreciar el sesgo que producen cada una de ellas, así como la variación que producen. Las técnicas que utilizadas han sido:*Holdout* $\frac{2}{3}/\frac{1}{3}$, *3 Repeticiones de Holdout* $\frac{2}{3}/\frac{1}{3}$, *Validación Cruzada de 10 capas* y *3 Repeticiones de Validación Cruzada de 10 capas*. Dichas metodologías experimentales se describirán en cada una de sus correspondientes secciones. A continuación se describen brevemente los algoritmos y conjuntos de datos utilizados para las labores experimentales.

Para la realización de los experimentos se ha utilizado la biblioteca **Weka**[too], que permite la realización de distintas tareas de entrenamiento así como verificación relacionadas con la *minería de datos* y los *algoritmos de aprendizaje automático* de manera sencilla.

1.1. Algoritmos

Los algoritmos utilizados para las tareas de aprendizaje pertenecen a la categoría de *Aprendizaje Inductivo Basado en el Error*. Ambos algoritmos se basan en *Aprendizaje Supervisado*, es decir, en la fase de entrenamiento utilizan el valor de la clase de destino como medida del error, el cual tratan de reducir al máximo. Mediante dicha estrategia tratan de conseguir clasificar correctamente las instancias futuras.

- **J48**: Es la implementación en Java de *C4.5*, un método de generación de árboles de decisión basado en la *Teoría de la Información*. En cada iteración trata de maximizar la ganancia de información producida tras cada partición con respecto de la clase de destino. Además, proporciona otras mejoras como *poda de ramas* para evitar el sobreajuste, el uso de *valores continuos* o el tratamiento de *valores desconocidos*.
- **JRIP**: Es la implementación en Java de *RIPPER*, un método de aprendizaje supervisado basado en reglas cuyas siglas significan "*Repeated Incremental Pruning to Produce Error Reduction*", lo que puede entenderse como la eliminación de reglas que se cumplen con pocas instancias para reducir el sobreajuste producido en la fase de aprendizaje, que genera todo el conjunto de reglas posibles a partir de una determinada heurística.

1.2. Conjuntos de Datos

Se han utilizado **3** conjuntos de datos en los experimentos realizados. Estos se describen brevemente a continuación:

- **Labor**[data]: Está formado por *57 instancias* formadas por *16 atributos* de los cuales, 8 de ellos son de tipo numérico mientras que el resto son de carácter nominal. La clase de destino puede tomar *2 valores* distintos. El conjunto de datos se corresponde con resultados de negociaciones industriales en Canadá.
- **Soybean**[datb]: Está formado por *683 instancias* formadas por *35 atributos*, todos ellos de carácter nominal. La clase de destino puede tomar *19 valores* distintos. El conjunto de datos se corresponde con instancias referidas a atributos de plantas y la clase de destino representa el tipo de planta.
- **Vote**[datc]: Está formado por *435 instancias* formadas por *16 atributos*, todos ellos de carácter nominal. La clase de destino puede tomar *2 valores* distintos. El conjunto de datos se refiere a resultados de encuestas a ciudadanos estadounidenses para tratar de predecir si votarán al partido demócrata o republicano.

En las siguientes secciones se describen los experimentos realizados así como los resultados obtenidos en cada caso junto con una discusión acerca de los mismos. Algo a destacar es el uso de distintas semillas para la tarea de particionamiento, las cuales se han indicado en las tablas de resultados según corresponda.

2. Realizar un experimento aplicando Holdout $\frac{2}{3}/\frac{1}{3}$

El método de *Holdout* consiste en el particionamiento del conjunto global de datos en 2 subconjuntos. Dicho método de experimentación requiere como entrada el porcentaje de datos que se utilizará para la tarea de entrenamiento, del cual se deriva el que se utilizará para test. En este caso se ha decidido utilizar $\frac{2}{3}$ del conjunto de datos para entrenamiento y $\frac{1}{3}$ para test. El método de selección que utiliza *Holdout* para seleccionar las instancias que formarán cada conjunto es la *selección aleatoria sin reemplazamiento.*

[TODO Hablar sobre resultados]

Holdout $2/3, 1/3$		
Datos	Algoritmo	Tasa de Error
		Semilla_1
Labor	J48	0,105263
	JRIP	0,105263
Soybean	J48	0,094828
	JRIP	0,086207
Vote	J48	0,027027
	JRIP	0,033784

Tabla 1

3. Realizar tres experimentos adicionales aplicando Holdout $\frac{2}{3}/\frac{1}{3}$, anotando la tasa de error de cada experimento

El método de *Holdout repetido* consiste en realizar las mismas tareas que el descrito anteriormente, pero en este caso realiza la misma tarea durante un determinado número de veces. La razón de ello es tratar de minimizar la varianza de la tasa de error promediando los resultados de cada una de las repeticiones. En este caso se ha decidido realizar *3 repeticiones* variando la semilla utilizada para cada uno de los experimentos de *Holdout*. El tamaño de las particiones, al igual que en el caso anterior, ha sido de $\frac{2}{3}$ para entrenamiento y $\frac{1}{3}$ para test.

[TODO Hablar sobre resultados]

Holdout 2/3, 1/3 Repetido				
Datos	Algoritmo	Tasa de Error		
		Semilla_2	Semilla_3	Semilla_4
Labor	J48	0,157895	0,315789	0,105263
	JRIP	0,157895	0,210526	0,105263
Soybean	J48	0,112069	0,107759	0,137931
	JRIP	0,077586	0,116379	0,073276
Vote	J48	0,081081	0,054054	0,060811
	JRIP	0,054054	0,047297	0,047297

Tabla 2

4. Sobre los resultados calculados en la sección 3 determinarla tasa de error, la varianza y el intervalo de confianza del 95 %

En este ejercicio se realizan las tareas de promediación así como del cálculo de la desviación típica y los intervalos de confianza correspondientes a los resultados del ejercicio 3.

En el caso del cálculo de la esperanza, se ha seguido la definición de la ecuación (1), la cual se define como la media aritmética del error.

$$e(h) = \frac{\sum_{i=1}^{k} e_i(h)}{k} \tag{1}$$

En el caso de la desviación típica, se ha utilizado la ecuación (2), que se corresponde con la definición de desviación típica muestral, pero utilizando la cuasi-varianza, la cual reduce el sesgo que se podría producir respecto del valor poblacional.

$$S_e(h) = \sqrt{\frac{\sum_{i=1}^{k}(e_i(h) - e(h))^2}{k-1}} \tag{2}$$

En el caso de los intervalos de confianza, se ha utilizado la ecuación (3), la cual se apoya en la distribución *T de Student*, así como la esperanza (1) como punto de equilibrio y la desviación típica (2) calculadas previamente.

$$[e(h) - t_{N,k-1} * \frac{S_e(h)}{\sqrt{k}}, e(h) + t_{N,k-1} * \frac{S_e(h)}{\sqrt{k}}] \tag{3}$$

[TODO Hablar sobre resultados]

Holdout 2/3, 1/3: Global				
Datos	Algoritmo	Tasa de Error	Desviación Estandar	Intervalos
Labor	J48	0,192982	0,109561	[0,008277, 0,377686]
	JRIP	0,157894	0,052631	[0,069165, 0,246622]
Soybean	J48	0,119253	0,016318	[0,091743, 0,146762]
	JRIP	0,089080	0,023739	[0,049059, 0,129100]
Vote	J48	0,065315	0,014065	[0,041603, 0,089026]
	JRIP	0,049549	0,003901	[0,042972, 0,056125]

Tabla 3

5. Realizar un experimento de validación cruzada de 10 particiones, calculando la tasa de error

[TODO]

Validación Cruzada		
Datos	Algoritmo	Tasa de Error
		Semilla_1
Labor	J48	0,263158
	JRIP	0,228070
Soybean	J48	0,084919
	JRIP	0,077599
Vote	J48	0,036782
	JRIP	0,045977

Tabla 4

6. Realizar tres experimentos de validación cruzada de 10 particiones, anotando la tasa de error

[TODO]

Validación Cruzada Repetida				
Datos	Algoritmo	Tasa de Error		
		Semilla_2	Semilla_3	Semilla_4
Labor	J48	0,263158	0,263158	0,245614
	JRIP	0,140351	0,157895	0,157895
Soybean	J48	0,098097	0,090776	0,079063
	JRIP	0,086384	0,068814	0,081991
Vote	J48	0,032184	0,036782	0,034483
	JRIP	0,043678	0,041379	0,03908

Tabla 5

7. Sobre los resultados calculados en la sección 6 determinarla tasa de error

[TODO]

Validación Cruzada Repetida: Global		
Datos	Algoritmo	Tasa de Error
Labor	J48	0,25731
	JRIP	0,152047
Soybean	J48	0,089312
	JRIP	0,079063
Vote	J48	0,034483
	JRIP	0,041379

Tabla 6

8. Conclusiones

[TODO]

Conjunto de Datos: Labor				
Algoritmo	Holdout	Holdout Repetido	Validación Cruzada	Validación Cruzada Repetida
J48	0,105263	0,192982	0,263158	0,25731
JRIP	0,105263	0,157894	0,228070	0,152047

Tabla 7

Conjunto de Datos: Soybean				
Algoritmo	Holdout	Holdout Repetido	Validación Cruzada	Validación Cruzada Repetida
J48	0,094828	0,119253	0,084919	0,089312
JRIP	0,086207	0,089080	0,077599	0,079063

Tabla 8

Conjunto de Datos: Vote				
Algoritmo	Holdout	Holdout Repetido	Validación Cruzada	Validación Cruzada Repetida
J48	0,027027	0,065315	0,036782	0,034483
JRIP	0,033784	0,049549	0,045977	0,041379

Tabla 9

Referencias

[CCAG17] Teodoro Calonge Cano and Carlos Javier Alonso GonzáLez. Técnicas de Aprendizaje Autómatico, 2016/17.

[data] Labor Data Set. `http://storm.cis.fordham.edu/~gweiss/data-mining/weka-data/labor.arff`.

[datb] Soybean Data Set. `http://storm.cis.fordham.edu/~gweiss/data-mining/weka-data/soybean.arff`.

[datc] Vote Data Set. `http://storm.cis.fordham.edu/~gweiss/data-mining/weka-data/vote.arff`.

[GP17] Sergio García Prado. Técnicas de aprendize automático: Evaluación de Hipótesis. `https://github.com/garciparedes/machine-learning-hypothesis-evaluation`, 2017.

[too] Weka. `http://www.cs.waikato.ac.nz/ml/weka/`.