Selección de características

Paloma Pérez G.
paperez@dcc.uchile.cl
Inteligencia artificial, EL4106 - Tarea 4

Resumen

Keywords – características, selección, entrenamiento, prueba

1. Introducción

Tanto problemas de estadística como de inteligencia artificial es posible encontrar dificultades al momento de determinar que información es útil o relevante para poder extraer conclusiones más precisas respecto del fenómeno que se esté estudiando. Esta información relevante o útil no es simple de detectar al observar los datos debido a que estos podrían tener una gran dimensionalidad o su influencia en los resultados no resulta ser del todo obvia. Para estos casos, en los cuales se busca construir un modelo de clasificacón sencillo pero a su vez lo más calibrado posible se debe explorar la información entregada a este y determinar que información es la más apropiada para tal acometido. Para este fin, es necesario hacer una selección de la información la que se traduce en la selección de características o atributos de las instancias de las clases que se quiere detectar.

En este trabajo se pretende realizar un acercamiento al proceso de selección de atributos usando dos diferentes tipos de bases de datos: una relacionada con la determinación de diabetes y la otra relacionada con la determinación de dígitos a partir de información obtenida de manuscritos. Este acercamiento se realizará usando el software WEKA.

1.1. Estructura del informe

Las siguientes secciones corresponden a: **Marco teórico**, en el cual se responde a preguntas planteadas respecto del la selección de características, de su proceso y que tipo métodos podrían apoyar en el proceso. Posteriormente, en la sección **??** se estudian los resultados obtenidos para las diferentes experiencias con el conjunto de datos de diabetes. En la sección **??** se describe la transformación del archivo CSV de los datos de dígitos a un archivo ARFF (para poder cargarlo en el software WEKA), y de la clasificación sin y con selección de atributos para estos datos. Luego en la sección **??** se tiene un análsis global de los resultados más importantes, que luego son resumidos en la sección de conclusiones **??**.

2. Marco teórico

2.1. Selección de características

En aprendizaje de máquinas y estadística, la *selección de características* corresponde esencialmente a la selección de un subconjunto de atributos relevantes para la construcción de un modelo. Esta selección se realiza por las siguientes razones:

1. Simplificación del modelo para lograr una interpretación más fácil de conseguir.
2. Disminuir los tiempos de entrenamiento
3. Para evitar la **maldición de la dimensionalidad** [1], la cual no sólo complejiza el entrenamiento, sino también exige una mayor cantidad de datos de entrenamiento.
4. Lograr una mejor generalización del modelo, reduciendo el overfitting (i.e. para reducir la varianza).

Su finalidad es la de deshacerse de información irrelevante o que no aporta información significativa dentro del modelo.

2.2. Etapas de la selección de características

1. **Generación de subconjuntos:** Es un procedimiento de sondeo que genera subconjuntos candidatos de características para la evaluación, basándose en una *estrategia de búsqueda* [2].
2. **Evaluación:** Cada subconjunto candidato es evaluado y comparado con los resultados obtenidos por el mejor subconjunto previo (el que ha obtenido mejores resultados en clasificación) de acuerdo a un *criterio de evaluación*. La generación de estos conjuntos dependerá de la métrica de evaluación, la cual puede corresponder a una de las siguientes categorías: de filtrado, wrappers o método embebido.
3. **Criterio de parada:**
4. **Validación:** El subconjunto seleccionado usualmente requiere de ser validado a través de conocimiento que se tenga de los datos o a través de diferentes pruebas usando datos sintéticos o reales.

	NEG	POS
NEG	161	17
POS	37	46

Cuadro 1: Matriz de confusión obtenida para datos de diabete usando todos los atributos.

	NEG	POS
NEG	160	18
POS	36	47

Cuadro 2: Matriz de confusión obtenida descartando el atributo *skin*.

2.3. Metodologías para la generación de subconjuntos

2.4. Metodologías de evaluación de subconjuntos

3. Selección de características y clasificación de datos de diabetes

En esta sección usamos un clasificador SMO con los valores prestablecidos (o por defecto) definidos en el software WEKA, con el cual entrenamos el 66 % de los datos, mientras que el resto es destinado a la evalución (conjunto de prueba).

3.1. Clasificación por defecto

La clasificación obtenida usando todas las características de los datos de diabetes entrega la siguiente matriz de confusión.

De estos resultados se desprende que se logra un 79.31 % de instancias correctamente clasificadas.

3.2. Clasificación seleccionando características usando BestFirst y WrapperSubsetEval

Para este experimento usamos el seleccionador de conjuntos BestFirst y el evaluador WrapperSubsetEval. Este último debe ser configurado para poder realizar las evaluaciones de los subconjuntos usando SMO. En esta ocasión, los resultados obtenidos por la selección de características indican que la información aportada por el feature *skin* no es relevante. Descartando este atributo se obtiene la siguiete matriz de confusión.

De estos resultados se desprende que se logra un 79.31 % de instancias correctamente clasificadas. Es decir, se mantiene el porcentaje anterior.

3.3. Clasificación seleccionando sólo 4 características usando Ranker e InfoGainAttributeEval

Para este experimento usamos el seleccionador de conjuntos Ranker y el evaluador InfoGainAttributeEval. En esta oportunidad se estima que los atributos *plas*, *insu*, *mass* y *age* son las cuatro primeras características que mejor describen el modelo aportando información significativa.

	NEG	POS
NEG	159	19
POS	33	50

Usando estas features se obtiene un 80.07 % de instancias correctamente clasificadas.

	NEG	POS
NEG	164	14
POS	36	47

Cuadro 3: Matriz de confusón usando los dos mejores atributos: *plas* y *mass*

.

3.4. Clasificación seleccionando sólo 2 características usando Ranker e InfoGainAttributeEval

En este experimento nuevamente usamos Ranker y el evaluador InfoGainAttributeEval, sin embargo esta vez escogemos las 2 mejores features, las que resultan ser *plass* y *mass*.

En esta oportunidad, el resultado de instancias correctamente clasificadasa mejoró sustancialmente, alcanzando un 80.84 %.

4. Selección de atributos y clasificación de dígitos

4.1. Obtención de archivo ARFF

La transformación del

Listing 1: csv2arff.py

```
1 def write_arff(filepath):
2     data=np.loadtxt(filepath, delimiter=',',dtype=int)
3     rows, cols = data.shape
4     arff_head = write_str(cols-1)
5     np.savetxt('digits.arff', data, delimiter=',', fmt='%d',header↩
        =arff_head)
6
7 def write_str(n_attrs):
8
9     rel = '@relation orh_digits\n'
10    str_tmp = '@attribute \'attr_%d\' numeric\n'
11    str_attrs = ''
12
13    for i in range(1, n_attrs+1):
14        str_attrs = str_attrs +str_tmp  % i
15
16    str_cl = '@attribute \'class\' { 0, 1, 2, 3, 4, 5, 6, 7, 8, 9 }\n'
17
18    return rel+str_attrs+str_cl+'@data\n'
19
20 if __name__ == "__main__":
21     write_arff('./1_digits.txt')
```

	0	1	2	3	4	5	6	7	8	9
0	178	0	0	0	0	0	0	0	0	0
1	0	176	0	0	1	0	0	0	1	2
2	0	0	189	1	0	0	1	0	0	0
3	0	0	0	198	0	0	0	0	0	1
4	0	0	0	0	203	0	0	0	0	0
5	0	0	0	0	0	190	0	0	0	0
6	0	0	0	0	2	0	216	0	0	0
7	0	0	0	1	0	0	0	171	0	0
8	0	2	0	1	1	1	1	1	177	2
9	0	0	0	2	1	0	0	0	1	190

Cuadro 4: Matriz de confusión obtenida para datos de diabete usando todos los atributos.

	0	1	2	3	4	5	6	7	8	9
0	178	0	0	0	0	0	0	0	0	0
1	0	175	1	0	0	0	0	0	2	2
2	1	0	190	0	0	0	0	0	0	0
3	0	0	0	197	0	1	0	0	0	1
4	1	1	0	0	199	0	0	0	0	2
5	1	0	0	0	0	189	0	0	0	0
6	0	0	0	0	3	0	215	0	0	0
7	1	0	0	2	0	0	0	169	0	0
8	0	8	1	1	1	1	0	1	171	2
9	0	0	0	2	0	0	0	0	1	191

Cuadro 5: Matriz de confusión obtenida para datos de diabete usando todos los atributos.

4.2. Clasificación por defecto

4.3. Clasificación seleccionando características con BestFirst y WrapperSubsetEval

5. Conclusión

Referencias

[1] R. Bellman, *Dynamic programming*. Courier Corporation, 2013.

[2] H. Liu and L. Yu, "Yu, l.: Toward integrating feature selection algorithm for classification and clustering. ieee transaction on knowledge and data engineering 17(4), 491-502," vol. 17, pp. 491–502, 04 2005.