

Árboles de Clasificación usando una Estimación Bayesiana

J. G. Castellano, S. Moral, A. Cano

Uncertainty Treatment in Artificial Intelligence Research Group
Department of Computer Science and Artificial Intelligence

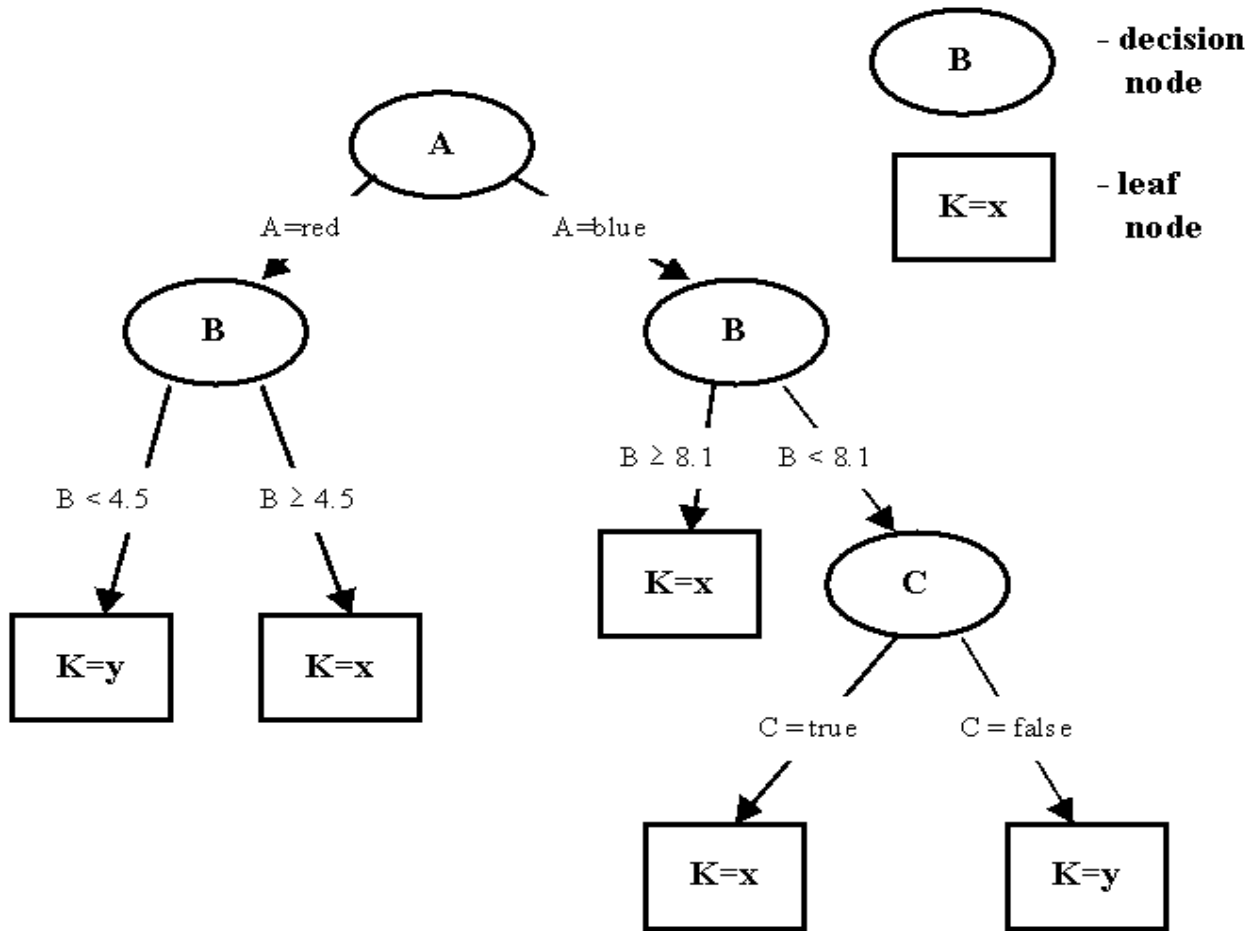
Contenido

- Introducción
- Árboles de clasificación
- Árboles de clasificación usando una estimación bayesiana
- Métodos de podas
- Problemas
- Experimentos (error de entrenamiento, error de test, número de nodos)
- Conclusiones

Introducción

- Los árboles de clasificación son clasificadores
- Representación del conocimiento
- Algoritmos clásicos: ID3 y C4.5
- Decisiones a tomar en la construcción del árbol
- Proceso de construcción
- La distribución de Dirichlet

Introduction



Arboles de clasificación

- C4.5 como evolución de ID3
- Para cada nodo: subconjunto de entrenamiento
- Seleccionar la mejor variables
- Expandir el árbol
- Criterio de parada

Arboles de clasificación 2

- Medida de la Información de Shannon

$$H(T) = \sum_{i=1}^n p_i \log_2(1/p_i)$$

- Probabilidad

$$H(X, T) = p(X = v_1 / T)H(T_1) + \dots$$

- Entropía

$$+ p(X = v_n / T)H(T_n)$$

- Entropía siguiente nivel

$$gain(X) = H(T) - H(X, T)$$

- Ganancia

- Información ruptura

$$SplitInfo(X, T) = \sum_{j=1}^n p(X = v_j / T) \log_2(1/p(x = v_j / T))$$

- Ratio de ganancia

$$GainRatio(X) = \frac{gain(x)}{SplitInfo(x)}$$

Arboles usando una distribución de Dirichlet

- Distribución de Dirichlet
- Parametros s, t
- Probabilidad usando Distribución Dirichlet
- Entropía y entropía siguiente nivel
- Nueva condición de parada

$$p(q) \propto \prod_j^k q_j^{st_j-1}$$

$$s = 2 \quad t = (1/n, \dots, 1/n)$$

$$p_j = \frac{s \cdot t_j + c_j}{s + m}$$

Metodos de poda

- Reduced Error Pruning
- Pessimistic Error Pruning
- Minimum Error Pruning
- Critical Value Pruning
- Cost-Complexity Pruning (CART)
- Error Based Pruning

Problemas

- Problemas de UCI usados por Acid y Abellán.
- Conjuntos de datos preprocesados
- Varios dominios

Data Set	Total	Train.	Test	Var.	Classes
Breast-cancer	286	184	93	9	2
Breast	699	457	266	10	2
Cleveland-nom	303	202	99	7	5
Cleveland	300	200	97	13	5
Pima	768	512	256	8	2
Heart	270	180	90	13	2
Hepatitis	155	53	27	19	2
Flare1	323	215	108	12	7
German	1000	670	330	24	2
Australian	690	460	230	14	2
Monks1	556	124	432	6	2
Monks2	601	169	432	6	2
Vote-irvine	435	290	145	16	2
soybean-small	47	31	16	35	4
soybean-large	683	374	188	35	19

Experimentos: Error de Entr.

Data Set	ID3	C4.5	Dirichlet	ID3+REP	C4.5+EBP	Dirichlet+REP
Breast-cancer	0.0217	0.0217	0.0706	0.3097	0.1847	0.1793
Breast	0.0131	0.0151	0.0153	0.0306	0.0328	0.0306
Cleveland-nom	0.2079	0.2128	0.2821	0.2178	0.4504	0.2970
Cleveland	0.0000	0.0000	0.1200	0.0000	0.4500	0.4500
Pima	0.1503	0.1601	0.1660	0.2187	0.2441	0.2050
Heart	0.0055	0.0055	0.0277	0.1388	0.2222	0.0888
Hepatitis	0.0000	0.0000	0.0188	0.1320	0.0754	0.0188
Flare1	0.0976	0.0976	0.1627	0.0976	0.1534	0.1627
German	0.0000	0.0000	0.0388	0.3059	0.3059	0.3059
Australian	0.0043	0.0043	0.0478	0.0586	0.1413	0.1000
Monks1	0.0161	0.0161	0.0000	0.1048	0.2661	0.1290
Monks2	0.0650	0.0591	0.0650	0.0650	0.3786	0.3786
Vote-irvine	0.0000	0.0000	0.0068	0.0379	0.0379	0.0379
soybean-small	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
soybean-large	0.0026	0.0026	0.0614	0.0026	0.0294	0.0614
Average Error	0.0389	0.0397	0.0722	0.1147	0.1981	0.1528

Experimentos: Error de test (I)

Data Set	ID3	C4.5	Dirichlet
Breast-cancer	0.4193	0.3870	0.2795
Breast	0.0353	0.0398	0.0353
Cleveland-nom	0.5252	0.5353	0.4949
Cleveland	0.5257	0.5051	0.4742
Pima	0.2929	0.2773	0.2773
Heart	0.1888	0.2777	0.2333
Hepatitis	0.1851	0.1851	0.1481
Flare1	0.3611	0.3611	0.3703
German	0.3424	0.3060	0.3303
Australian	0.2000	0.2043	0.1739
Monks1	0.1342	0.1527	0.0833
Monks2	0.3333	0.3217	0.3263
Vote-irvine	0.0620	0.0482	0.0482
soybean-small	0.0000	0.0000	0.0000
soybean-large	0.1489	0.0638	0.1063
Average Error	0.2503	0.2443	0.2254

Experimentos: Error de Test (II)

Data Set	ID3+REP	C4.5+EBP	Dirich.+REP
Breast-cancer	0.2580	0.2688	0.2365
Breast	0.0486	0.0398	0.0398
Cleveland-nom	0.5151	0.4747	0.4949
Cleveland	0.5257	0.4845	0.4845
Pima	0.2578	0.2187	0.2656
Heart	0.2333	0.2222	0.2444
Hepatitis	0.2222	0.1851	0.1481
Flare1	0.3611	0.3703	0.3703
German	0.2878	0.2878	0.2878
Australian	0.1565	0.1521	0.1695
Monks1	0.1620	0.2500	0.1388
Monks2	0.3333	0.3287	0.3287
Vote-irvine	0.0482	0.0482	0.0482
soybean-small	0.0000	0.0000	0.0000
soybean-large	0.1489	0.0585	0.1063
Average Error	0.2372	0.2259	0.2102

Experimentation: Nodes Number (I)

Data Set	ID3	C4.5	Dirichlet
Breast-cancer	342	270	153
Breast	59	63	33
Cleveland-nom	160	152	60
Cleveland	193	305	106
Pima	273	247	137
Heart	110	116	62
Hepatitis	11	17	7
Flare1	301	234	34
German	493	511	303
Australian	383	406	165
Monks1	88	87	79
Monks2	146	142	118
Vote-irvine	46	46	28
soybean-small	6	8	6
soybean-large	220	183	66
Average Error	188.73	185.80	90.46

Experimentos: Número de Nodos (II)

Data Set	ID3+REP	C4.5+EBP	Dirich.+REP
Breast-cancer	1	55	48
Breast	21	15	17
Cleveland-nom	150	1	41
Cleveland	193	1	1
Pima	53	5	47
Heart	37	10	37
Hepatitis	1	5	7
Flare1	301	38	34
German	1	1	1
Australian	156	3	18
Monks1	53	5	28
Monks2	146	1	1
Vote-irvine	7	7	7
soybean-small	6	8	6
soybean-large	220	85	66
Average Error	89.73	16.00	22.43

Experimentos: Métodos de Poda

Data Set	ID3	C4.5	Dirich.
Train. Error(unpruned)	0.0389	0.0397	0.0722
Train. Error with REP	0.1147	0.0861	0.1528
Train. Error with EBP	0.2111	0.1981	0.1956
Test Error(unpruned)	0.2503	0.2443	0.2254
Test Error with REP	0.2372	0.2383	0.2102
Test Error with EBP	0.2375	0.2259	0.2196

conclusiones

- Nuevo método para construir árboles de clasificación
- Más Simple (entropía vs gain ratio)
- Genera árboles más pequeños (y es más simple) => más rápido
- Mejor error de test
- Peor error de entrenamiento => menor sobre ajuste
- Menos nodos en la construcción del árbol

Trabajo Futuro

- Tratar datos continuos
- Aplicación selectiva de la distribución de Dirichlet
- Probar más métodos de poda
- Aplicaciones para éste método:
Imputación de datos perdidos