

Practical Application 2

Machine Learning

Jan Cerezo Pomykol
j.cerezo@alumnos.upm.es

Universidad Politécnica de Madrid
ETSIINF

November 22, 2022

Problem Description

Dry Bean Dataset:

- 13611 instances
- 16 variables
- 7 classes
- [Source](#)



- Area
- Perimeter
- Major axis length
- Minor axis length
- Aspect ratio
- Eccentricity
- Convex area
- Equivalent diameter
- Extent
- Solidity
- Roundness
- Compactness
- ShapeFactor1
- ShapeFactor2
- ShapeFactor3
- ShapeFactor4
- Seker
- Barbunya
- Bombay
- Cali
- Dermosan
- Horoz
- Sira

- Software: **Weka**
- Classification algorithms:
 - Logistic Regression
 - Naive Bayes
 - Tree Augmented Naive Bayes
 - Linear Discriminant Analysis
 - Fusion
 - Stacking
 - Bagging
 - Random Forest
 - Boosting
 - Naive Bayes Tree
 - Logistic Model Tress
- Feature Subset Selection
 - No FSS
 - Univariate Filter
 - Multivariate Filter
 - Wrapper Approach

Algorithm	Weka Function
Logistic Regression	functions.Logistic
Naive Bayes	bayes.NaiveBayes
Tree Augmented Naive Bayes	bayes.BayesNet
Linear Discriminant Analysis	functions.LDA
Fusion	meta.Vote
Stacking	meta.Stacking
Bagging	meta.Bagging
Random Forest	trees.RandomForest
Boosting	meta.AdaBoostM1
Naive Bayes Tree	trees.NBTtree
Logistic Model Trees	trees.LMT

FSS algorithm	Weka Function
No FSS	-
Univariate Filter	attributeSelection.InfoGainAttributeEval
Multivariate Filter	attributeSelection.CfsSubsetEval
Wrapper Approach	attributeSelection.WrapperSubsetEval

Results

Selected Attributes

Attribute	No FSS	Univariate	Multivariate	Wrapper (Logistic)	Wrapper (Naive Bayes)	Wrapper (TAN)	Wrapper (LDA)	Wrapper (Fusion)	Wrapper (Stacking)	Wrapper (Bagging)	Wrapper (Random Forest)	Wrapper (Boosting)	Wrapper (NB Tree)	Wrapper (LMT)
Area		•					•			•	•	•		•
Perimeter	•	•	•	•	•	•	•	•	•	•		•	•	•
MajorAxisLength	•	•	•	•					•		•			•
MinorAxisLength	•	•	•	•		•		•	•				•	•
AspectRatio	•		•			•			•					
Eccentricity	•							•	•	•				
ConvexArea	•	•	•	•			•	•	•	•				•
EquivDiameter	•	•		•				•	•	•				•
Extent	•		•	•		•	•	•	•	•	•	•		•
Solidity	•			•				•	•	•	•		•	•
Roundness	•		•	•	•	•		•	•	•	•		•	•
Compactness	•		•	•		•	•	•	•	•	•		•	•
ShapeFactor1	•	•	•	•	•	•		•		•	•		•	
ShapeFactor2	•	•	•	•					•	•				•
ShapeFactor3	•										•	•		
ShapeFactor4	•		•	•	•	•	•	•	•	•	•		•	•
N attributes	16	8	11	11	5	8	6	7	10	9	9	4	6	9

Results

Classifier scores

Dataset	Logistic	Naive Bayes	TAN	LDA	Fusion	Stacking	Bagging	Random Forest	Boosting	NBTree	LMT
Original	92.60	89.71	91.47	90.18	91.26	91.28	89.72	92.52	89.71	89.57	92.49
Uni. Filter	92.14	84.09	89.90	89.22	90.00	90.29	84.03	91.04	84.09	87.67	91.94
Mult. Filter	92.57	90.20	91.24	90.05	91.58	91.74	90.31	92.47	90.20	90.63	92.41
Wr. (Logistic)	92.70	89.01	91.47	90.03	91.47	91.53	89.08	92.53	89.01	89.66	89.01
Wr. (N. Bayes)	92.09	91.23	91.54	89.84	89.01	91.77	91.21	92.16	91.23	91.55	92.16
Wr. (TAN)	92.36	90.76	91.60	89.83	91.72	91.62	90.80	92.33	90.76	90.69	92.27
Wr. (LDA)	92.30	88.23	90.42	91.17	91.35	91.56	88.34	91.74	88.23	89.57	92.35
Wr. (Fusion)	92.39	91.05	91.29	90.58	91.91	91.24	91.05	92.68	91.05	90.88	92.44
Wr. (Stacking)	92.55	89.42	91.66	89.86	91.38	92.20	89.45	92.46	89.42	89.97	92.41
Wr. (Bagging)	92.53	90.77	91.44	89.45	91.58	91.78	90.75	92.70	90.77	90.90	92.54
Wr. (R. Forest)	92.38	90.66	91.27	89.72	91.58	91.53	90.65	92.84	90.66	90.33	92.56
Wr. (Boosting)	91.12	80.83	89.60	88.85	89.39	89.69	80.89	91.11	80.83	89.48	91.42
Wr. (NBTree)	92.21	91.22	91.24	90.56	91.89	91.17	91.22	92.46	91.22	91.27	92.27
Wr. (LMT)	92.45	84.75	91.02	90.32	91.22	91.30	84.80	92.28	84.75	89.82	92.52

Results

Training time

Dataset	Logistic	Naive Bayes	TAN	LDA	Fusion	Stacking	Bagging	Random Forest	Boosting	NBTree	LMT
Original	57.5	0.02	0.14	0.02	0.17	1.67	0.19	3.99	1.17	11.51	14.38
Uni. Filter	2.39	0.01	0.06	0.01	0.07	0.7	0.09	3.1	0.61	5.83	10.58
Mult. Filter	3.89	0.01	0.09	0.01	0.11	1.06	0.15	3.18	0.84	7.17	11.41
Wr. (Logistic)	5.81	0.01	0.09	0.01	0.11	1.06	0.13	3.21	1.14	10.54	12.27
Wr. (N. Bayes)	1.37	0.01	0.03	0.01	0.04	0.45	0.08	2.38	0.46	1.92	8.39
Wr. (TAN)	2.18	0.01	0.06	0.01	0.07	0.74	0.1	3.14	0.51	6.52	9.58
Wr. (LDA)	1.99	0.01	0.04	0.01	0.07	0.53	0.08	2.48	0.39	2.9	13
Wr. (Fusion)	1.97	0.01	0.05	0.01	0.06	0.63	0.09	2.46	0.58	5.23	9.86
Wr. (Stacking)	2.69	0.01	0.07	0.01	0.09	0.92	0.12	3.21	0.99	8.71	10.74
Wr. (Bagging)	2.44	0.01	0.07	0.01	0.09	0.82	0.12	3.17	0.67	5.63	10.57
Wr. (R. Forest)	2.51	0.01	0.07	0.01	0.09	0.82	0.11	3.21	0.87	8.61	10.33
Wr. (Boosting)	0.97	0.01	0.03	0.01	0.03	0.36	0.06	2.18	0.46	1.43	8.26
Wr. (NBTree)	0.96	0.01	0.04	0.01	0.06	0.53	0.09	2.44	0.45	3.96	9.05
Wr. (LMT)	7.95	0.01	0.06	0.01	0.08	0.83	0.11	3.28	0.56	8.26	16.41

Results

Logistic Regression

Coefficients...

Variable	Class SEKER	BARBUNYA	BOMBAY	CALI	HOROZ	SIRA
Perimeter	122.6821	83.2646	722.999	57.7438	147.5876	-96.0632
roundness	5.1908	-17.4808	358.6025	16.7366	16.8368	-23.7141
Compactness	47.5217	10.6809	195.6578	-36.1733	-13.6638	-38.3752
ShapeFactor1	14.7379	-59.4434	348.0428	-95.3347	47.0252	-106.6375
ShapeFactor4	29.8703	8.9363	-134.3783	-16.5178	-17.6388	-10.7245
Intercept	-86.6715	11.6021	-642.3413	50.4722	-45.2959	121.7997

Odds Ratios...

Variable	Class SEKER	BARBUNYA	BOMBAY	CALI	HOROZ	SIRA
Perimeter	1.9060785863351933E53	1.4499412945562452E36	Infinity	1.1961829114581766E25	1.2487633551787505E64	0
roundness	179.6034	0	5.483927383541303E155	18561985.0514	20518650.4473	0
Compactness	4.349342686745699E20	43517.6216	9.399751905032624E84	0	0	0
ShapeFactor1	2515213.6001	0	1.422496489365307E151	0	2.6471799061105616E20	0
ShapeFactor4	9.386358439569479E12	7602.8993	0	0	0	0

Results

Naive Bayes

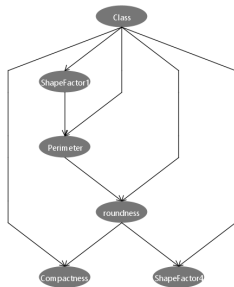
Attribute	Class						
	SEKER (0.15)	BARBUNYA (0.1)	BOMBAY (0.04)	CALI (0.12)	HOROS (0.14)	SIRA (0.19)	DERMASON (0.26)
=====							
Perimeter							
mean	0.1389	0.3569	0.7263	0.3648	0.2705	0.186	0.0962
roundness							
mean	0.9078	0.6198	0.748	0.7111	0.6083	0.7884	0.8352
Compactness							
mean	0.7391	0.4742	0.4385	0.3349	0.1739	0.4521	0.5149
ShapeFactor1							
mean	0.4635	0.3362	0.0865	0.3494	0.5511	0.5137	0.6486
ShapeFactor4							
mean	0.9741	0.9233	0.8484	0.8242	0.85	0.9165	0.9458

=== Confusion Matrix ===

a	b	c	d	e	f	g	<-- classified as
1922	27	0	0	1	55	22	a = SEKER
5	1161	1	106	12	37	0	b = BARBUNYA
0	0	522	0	0	0	0	c = BOMBAY
2	90	0	1488	39	11	0	d = CALI
0	6	0	28	1851	29	14	e = HOROS
47	7	0	10	67	2307	198	f = SIRA
87	4	0	0	23	265	3167	g = DERMASON

Results

Tree Augmented Naive Bayes



Probability Distribution Table For Class

SEKER	BARBUN...	BOMBAY	CALI	HOROZ	SIRA	DERMAS...
0,149	0,097	0,038	0,12	0,142	0,194	0,26

Probability Distribution Table For ShapeFactor1

Class	'(-inf...	'(0,17...	'(0,26...	'(0,32...	'(0,36...	'(0,38...	'(0,40...	'(0,43...	'(0,47...	'(0,51...	'(0,52...	'(0,53...	'(0,55...	'(0,59...	'(0,62...	'(0,66...	'(0,70...	'(0,76...
SEKER	0	0	0,001	0,015	0,026	0,064	0,128	0,306	0,347	0,062	0,023	0,021	0,004	0,002	0	0	0	0
BARBUN...	0,001	0,1	0,329	0,3	0,073	0,074	0,06	0,044	0,012	0	0,003	0	0,001	0	0	0	0	0
BOMBAY	0,984	0,001	0,001	0,001	0,001	0,001	0,001	0,001	0,001	0,001	0,001	0,001	0,001	0,001	0,001	0,001	0,001	0,001
CALI	0	0,024	0,233	0,407	0,146	0,097	0,058	0,023	0,009	0	0	0	0,001	0	0	0	0	0
HOROZ	0	0	0	0,002	0,001	0,002	0,014	0,081	0,199	0,088	0,067	0,116	0,171	0,128	0,063	0,055	0,012	0
SIRA	0	0	0	0	0,001	0,001	0,012	0,135	0,372	0,128	0,106	0,107	0,091	0,035	0,01	0,001	0	0
DERMASON	0	0	0	0	0	0	0	0	0,018	0,017	0,028	0,059	0,134	0,163	0,16	0,192	0,15	0,077

Practical Application 2

Machine Learning

Jan Cerezo Pomykol
`j.cerezo@alumnos.upm.es`

Universidad Politécnica de Madrid
ETSIINF

November 22, 2022