Interpretable Models to Predict Breast Cancer

Pedro Ferreira, MSc[1]; Inês Dutra, PhD[1,2];
Rogerio Salvini, PhD[3]; Elizabeth Burnside, MD, MPH, MS[4].

[1]CRACS-INESC TEC, Porto, Portugal
[2]DCC-FC, University of Porto, Portugal
[4]University of Wisconsin, Madison, USA
[3]Institute of Informatics, Federal University of Goiás, Brazil

Outline

- Breast Cancer
- Approach & Objectives
- Variables Relevance
- ILP vs SVM
- Interpretable Classifiers
- Malignant Rules
- Conclusions & Future Work

Outline

- Breast Cancer
- Approach & Objectives
- Variables Relevance
- ILP vs SVM
- Interpretable Classifiers
- Malignant Rules
- Conclusions & Future Work

Breast Cancer

Source: U.S. Breast Cancer Statistics [1] – accessed December 2016

Breast Cancer

Source: U.S. Breast Cancer Statistics [1] – accessed December 2016

Breast Cancer

Source: U.S. Breast Cancer Statistics [1] – accessed December 2016

Approach

Approach

- Several works in the **literature** use **propositional** ("**black box**") **approaches** to generate **prediction models**.

- In this work **we employ** the **Inductive Logic Programming** technique, whose prediction model is **based on first order rules**, to the domain of breast cancer.

(+) Interpretable Rules

ILP

Objectives

- **Generate** more **interpretable models** based on first-order logic

ILP

- **Compare ILP** performance results **with propositional classifiers**

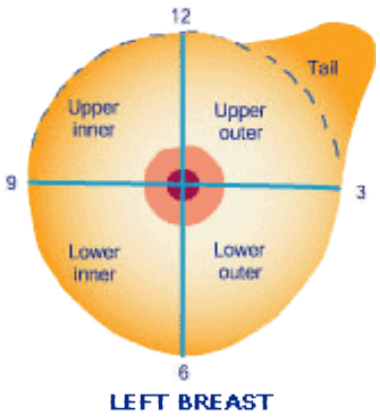
ILP vs SVM vs DT

- **Explore relevance** of some **variables** usually collected to predict breast cancer

Variables Relevance

MammoClass[2]

Classification of a mammogram based in a set of mammography findings



Enter Data

Patient's age	22
Mass size	10
Breast Composition	Almost entirely fat
Mass shape	Lobular
Mass clockface location	2.0
Mass margins (1)	Circumscribed
Mass margins (2)	
Mass margins worst	Mass Margins (1)
Mass density	Iso/Low
Side	Left
Quadrant	Lower Inner
Depth	Anterior

Predict Reset

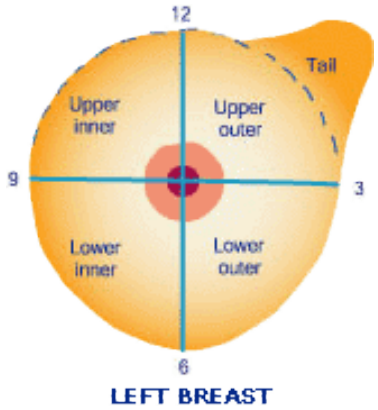
Result

Predicted mass density: iso (98.4%)

Prediction: mass benign with a probability of 99.9%.

Variables Relevance

- **Side**, **Depth**, **Clockface** and **Quadrant** are considered to be **non** indicative of **malignancy by expert** radiologists
- **<u>However</u>** some **studies show** that for some populations there can be a **prevalence of breast cancer** according to the **value** of some of these **variables**



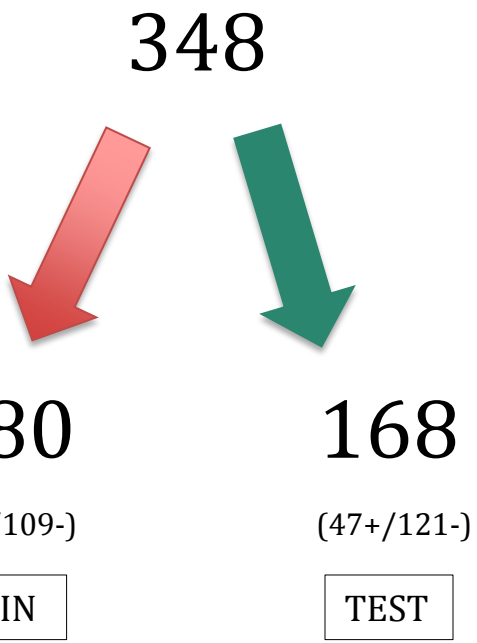
GEC-ESTRO [3] says that the **upper outer quadrant** is the most common site of origin of **breast cancer**

GEC-ESTRO [3] says that **breast cancer** is **more common in the left** than in the right breast. Other studies on laterality confirm this tendency [4]

Can we **remove** these **variables** and still obtain the same results with the **test set** in this sample?

Dataset

- Breast Masses
- Annotated Data
- Test independent from training set



[2] Ferreira, P., Fonseca, N.A., Dutra, I., Woods, R., Burnside, E.:

Predicting Malignancy from Mammography Findings and Image-Guided Core Biopsies.

In: Int. Journal of Data Mining and Bioinformatics, 2015.

Tools

ALEPH

- **ILP System**
- Written in **Prolog**
- **Powerful representation language**
- User may choose the order of generation of rules, **change** the **evaluation function** and the search order
- **Open Source**

WEKA

- Set of **machine learning algorithms** for data mining tasks
- Written in **Java**
- Contains tools for **data pre-processing**, **classification**, regression, clustering, **association rules**, etc
- Well-suited for developing new machine learning schemes
- **Free software**

Methodology – Experiments

- **A –** Trains **SVM** on 180, **without the 4 variables**, and evaluates on 168 test set
- **Prev**[2] **–** Trains **SVM** on 180, **using all variables**, and evaluates on 168 test set

- **B1 –** Trains **Aleph** on 180, **using all variables**, and evaluates on 168 test set
- **B2 –** Trains **Aleph** on 180, **without the 4 variables**, and evaluates on 168 test set

Variables Relevance - Results

TABLE I
PERFORMANCE OF CLASSIFIERS ON TEST SET

	Platform	Exp.	CCI	K	F	AUROC	TPR	P	TNR
All vars.	Aleph	* B1	77.4	0.37	0.52	–	0.43	0.65	0.91
w/o 4 vars.	Aleph	* B2	79.8	0.41	0.52	–	0.40	0.76	0.95
All vars.	WEKA	** Prev	79.2	0.47	0.62	0.82	0.60	0.64	0.87
w/o 4 vars.	WEKA	A	81.0	0.51	0.64	0.85	0.60	0.68	0.89

McNemar's Tests

B1 vs B2 -> p-value = 0.18
Prev vs A -> p-value = 0.55

Not Statistically Significant

4 VARS. NOT RELEVANT THIS DATASET

B1 vs Prev -> p-value = 0.02

Statistically Significant

∗ *noise* = 0 | *evalfn* = *coverage*

∗∗ results published in [2]

Variables Relevance - Results

TABLE I
PERFORMANCE OF CLASSIFIERS ON TEST SET

	Platform	Exp.	CCI	K	F	AUROC	TPR	P	TNR
All vars.	Aleph	*B1	77.4	0.37	0.52	–	0.43	0.65	0.91
w/o 4 vars.	Aleph	*B2	79.8	0.41	0.52	–	0.40	0.76	0.95
All vars.	WEKA	**Prev	79.2	0.47	0.62	0.82	0.60	0.64	0.87
w/o 4 vars.	WEKA	A	81.0	0.51	0.64	0.85	0.60	0.68	0.89

McNemar's Tests

B1 vs B2 -> p-value = 0.18
Prev vs A -> p-value = 0.55

Not Statistically Significant

B1 vs Prev -> p-value = 0.02

Statistically Significant

4 VARS. NOT RELEVANT THIS DATASET

ILP WORSE THAN SVM

ILP vs SVM

Searching for ILP classifiers
that can be better than the SVM…

Aleph's Internal Parameters

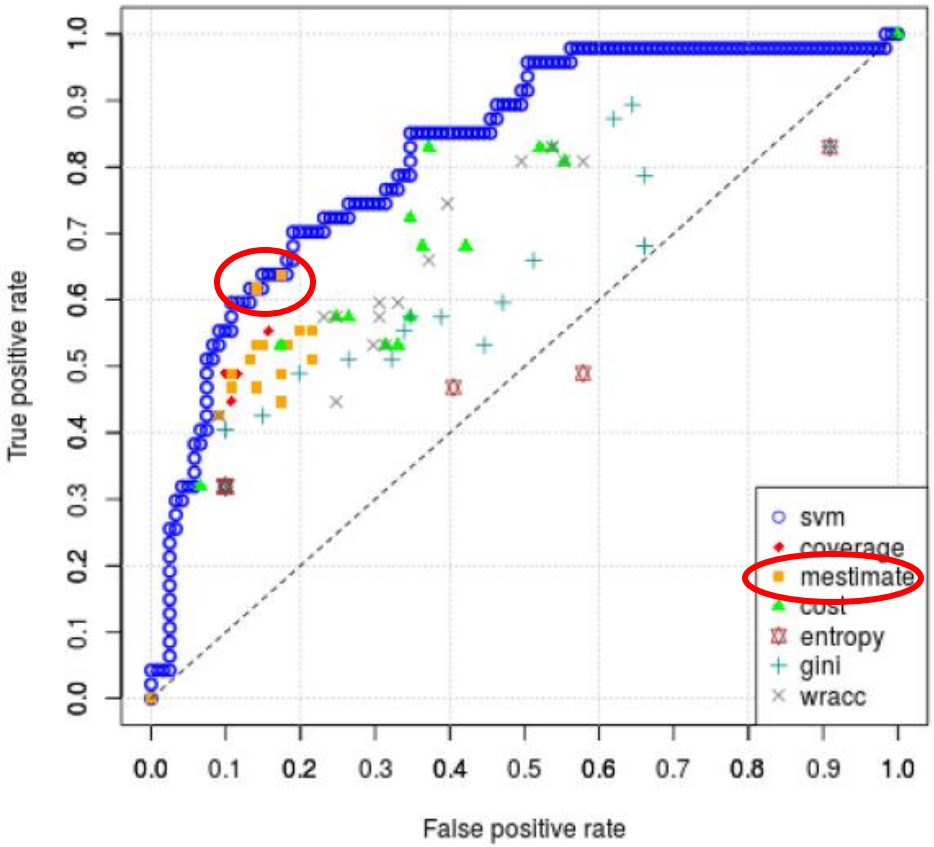
- ***Noise* –** controls the maximum number of **false positives allowed** by the model during training

- ***Evalfn* –** controls the **evaluation function** used to assess the quality of each hypothesis generated

  - *coverage,* ***mestimate****, cost, entropy, gini, and wracc*

ILP vs SVM

Fig. 1. ROC points for SVM and ILP



McNemar's Tests

noise **= 19 -> p-value = 0.84**
noise **= 93 -> p-value = 0.23**

<u>Not</u> Statistically Significant

ILP
EQUAL TO
SVM

Interpretable Classifiers

Interpretable Classifiers

```
is_malignant(A) :-
        shape(A,'Round'),
        depth(A,'Middle'),
        density(A,high).
```

	TRAINING SET	TEST SET
Pos Cover by Rules	6	1
Neg Cover by Rules	0	0
TOTAL Pos /Negs	71 + / 109 -	47 + / 121 -

```
is_malignant(A) :-
        shape(A,'Irregular'),
        margins(A,'Spiculated').
```

	TRAINING SET	TEST SET
Pos Cover by Rules	17	7
Neg Cover by Rules	0	0
TOTAL Pos /Negs	71 + / 109 -	47 + / 121 -

Malignant Rules

Malignant Rules

TRAIN

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1		Malignant Rules								Benign Rules							
2	Instance	R1	R2	R3	R4	R5	R6	R7	R8	R1	R2	R3	R4	R5	R6	R7	R8
3	2	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
4	3	0	1	0	1	0	0	0	1	0	0	0	1	0	0	0	0
5	7	0	0	1	0	0	1	0	1	0	0	0	1	0	0	0	0
6	13	0	1	0	0	0	1	0	1	0	0	0	0	0	0	1	0
7	17	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0
8	23	0	0	0	1	1	0	0	0	1	0	0	1	0	1	0	1
9	24	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0
10	25	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
11	34	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
12	39	0	0	0	0	0	1	0	0	0	0	1	0	1	0	0	0
13	41	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
14	52	1	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0
15	59	0	0	1	0	1	0	0	0	0	0	0	0	0	0	1	0
16	60	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0
17	65	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1
18	68	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0
19	69	1	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0
20	74	1	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0
173	316	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
174	321	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
175	328	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0
176	332	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0
177	334	1	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0
178	336	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
179	338	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
180	339	0	0	0	0	1	0	1	0	1	0	0	0	0	0	0	0
181	342	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
182	348	0	1	0	0	0	0	0	0	1	0	0	0	0	0	1	0
183	**TP**	23	19	24	20	23	19	3	15	52	43	26	21	17	15	15	20
184	**TN**	103	101	108	97	103	99	101	102	63	61	65	59	65	69	61	60
185	**FP**	6	8	1	12	6	10	8	7	8	10	6	12	6	2	10	11
186	**FN**	48	52	47	51	48	52	68	56	57	66	83	88	92	94	94	89
187	**CCI**	**0,70**	**0,67**	**0,73**	**0,65**	**0,70**	**0,66**	**0,58**	**0,65**	**0,64**	**0,58**	**0,51**	**0,44**	**0,46**	**0,47**	**0,42**	**0,44**
188																	

train | test | predictions

Find | Find All | Search Formatted Display String | Match Case

Sheet 1 of 3 | Default

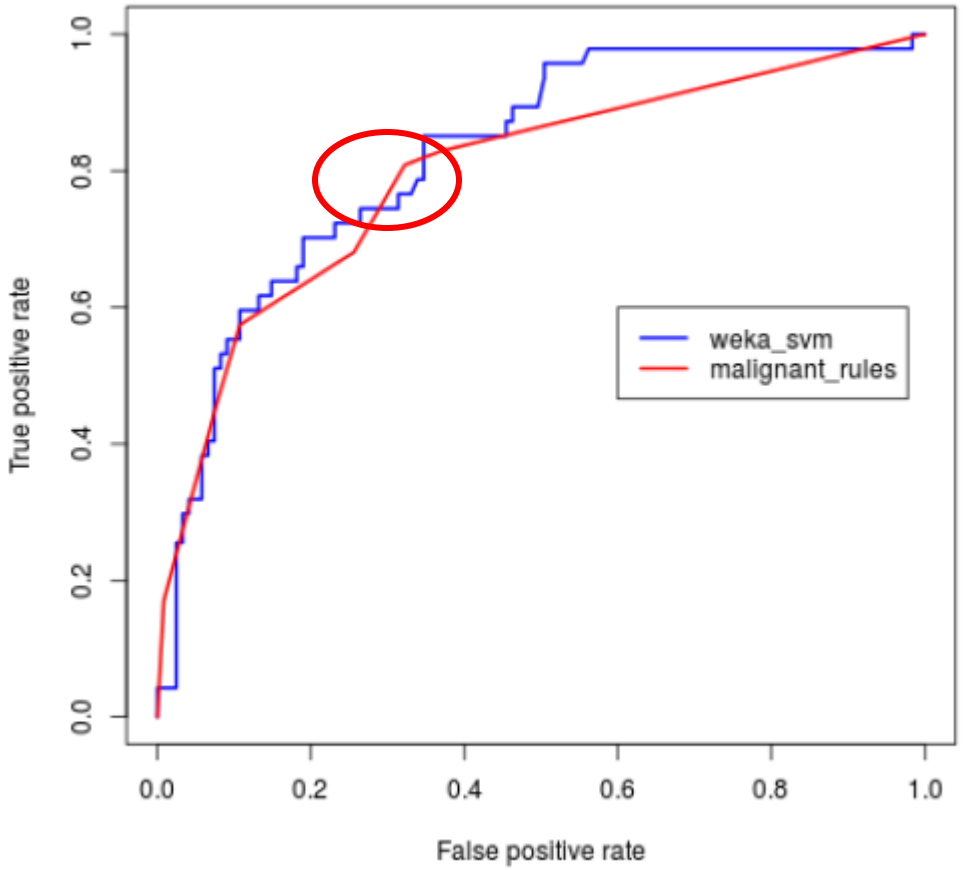
Malignant Rules

TEST

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	T	U
1			Malignant Predictions								Benign Predictions								Max Value	
2	Instance	Class	R1	R2	R3	R4	R5	R6	R7	R8	R1	R2	R3	R4	R5	R6	R7	R8	Malignant	Benign	All
3	4	1	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
4	5	1	0,00	0,67	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,46	0,00	0,42	0,00	0,67	0,46	0,67
5	6	1	0,00	0,00	0,00	0,00	0,70	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,70	0,00	0,70
6	10	1	0,70	0,00	0,00	0,00	0,70	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,70	0,00	0,70
7	12	1	0,00	0,00	0,00	0,65	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,42	0,44	0,65	0,44	0,65
8	20	1	0,00	0,00	0,00	0,65	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,65	0,00	0,65
9	32	1	0,70	0,00	0,00	0,00	0,00	0,66	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,47	0,00	0,00	0,70	0,47	0,70
10	33	1	0,00	0,67	0,73	0,00	0,00	0,66	0,00	0,65	0,00	0,00	0,00	0,00	0,00	0,00	0,42	0,00	0,73	0,42	0,73
11	37	1	0,70	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,51	0,00	0,00	0,00	0,42	0,00	0,70	0,51	0,70
12	40	1	0,00	0,00	0,00	0,65	0,00	0,00	0,00	0,00	0,00	0,00	0,51	0,00	0,00	0,00	0,00	0,00	0,65	0,51	0,65
13	63	1	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
14	66	1	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
15	70	1	0,00	0,67	0,00	0,00	0,00	0,00	0,00	0,65	0,00	0,00	0,00	0,00	0,00	0,00	0,42	0,00	0,67	0,42	0,67
16	71	1	0,00	0,00	0,00	0,00	0,70	0,66	0,00	0,00	0,00	0,00	0,00	0,44	0,00	0,00	0,00	0,00	0,70	0,44	0,70
17	72	1	0,00	0,00	0,00	0,00	0,70	0,66	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,70	0,00	0,70
18	80	1	0,70	0,00	0,73	0,00	0,00	0,66	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,73	0,00	0,73
19	90	1	0,00	0,00	0,00	0,65	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,44	0,46	0,00	0,00	0,00	0,65	0,46	0,65
20	91	1	0,00	0,67	0,00	0,65	0,00	0,00	0,00	0,65	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,67	0,00	0,67
21	114	1	0,00	0,00	0,00	0,00	0,00	0,66	0,00	0,00	0,64	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,66	0,64	0,66
22	117	1	0,70	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,58	0,51	0,00	0,46	0,00	0,00	0,00	0,70	0,58	0,70
23	122	1	0,00	0,00	0,73	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,73	0,00	0,73
24	129	1	0,70	0,00	0,73	0,65	0,00	0,00	0,00	0,00	0,00	0,00	0,51	0,00	0,00	0,00	0,00	0,00	0,73	0,51	0,73
25	141	1	0,00	0,00	0,73	0,00	0,00	0,00	0,00	0,65	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,73	0,00	0,73
26	154	1	0,00	0,67	0,73	0,65	0,00	0,00	0,58	0,65	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,73	0,00	0,73
27	160	1	0,00	0,00	0,00	0,00	0,00	0,66	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,66	0,00	0,66
28	164	1	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,58	0,00	0,00	0,00	0,00	0,00	0,44	0,00	0,58	0,58
29	166	1	0,00	0,00	0,73	0,00	0,00	0,00	0,00	0,65	0,00	0,58	0,00	0,00	0,00	0,00	0,42	0,00	0,73	0,58	0,73
30	181	1	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,64	0,00	0,00	0,00	0,00	0,00	0,00	0,44	0,00	0,64	0,64
31	182	1	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,58	0,00	0,44	0,46	0,00	0,00	0,00	0,00	0,58	0,58
32	197	1	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,47	0,00	0,44	0,00	0,47	0,47
33	203	1	0,00	0,00	0,00	0,00	0,70	0,00	0,00	0,00	0,00	0,00	0,00	0,44	0,00	0,00	0,00	0,44	0,70	0,44	0,70
34	204	1	0,70	0,00	0,00	0,00	0,70	0,66	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,70	0,00	0,70
35	215	1	0,00	0,00	0,00	0,00	0,00	0,66	0,00	0,00	0,00	0,00	0,00	0,00	0,46	0,00	0,00	0,44	0,66	0,46	0,66

Malignant Rules

Fig. 2. ROC points for SVM and malignant rules from ILP



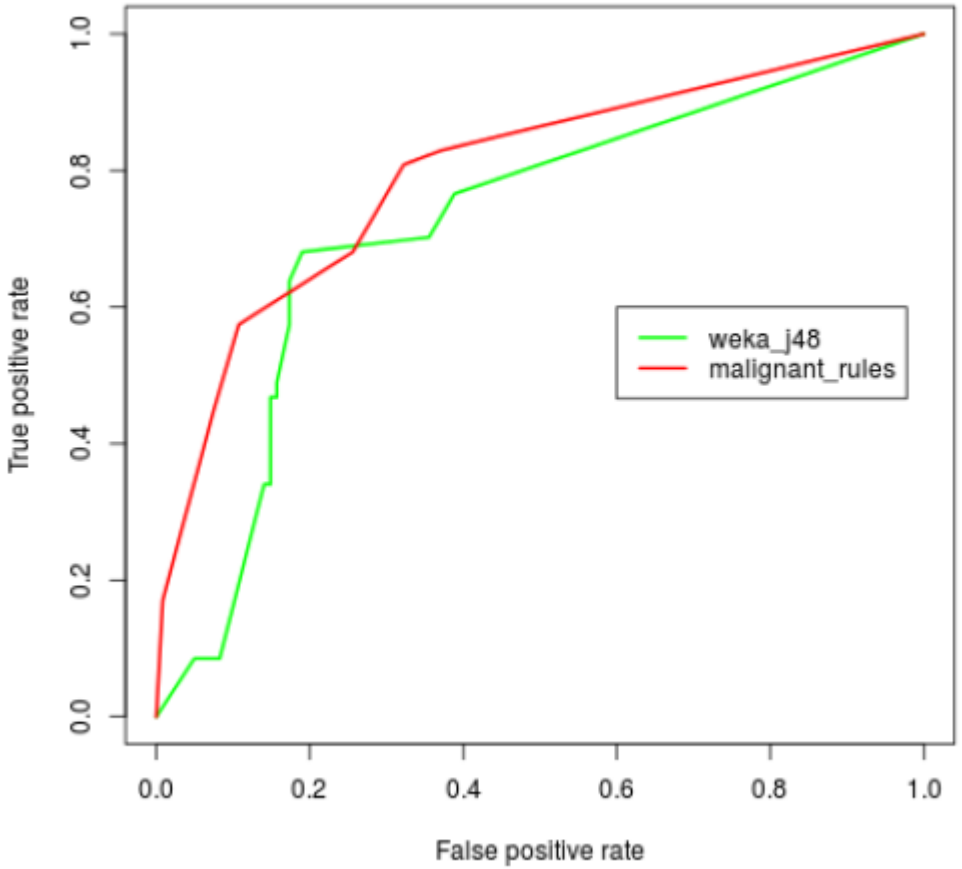
ILP

BETTER* THAN

SVM

Malignant Rules

Fig. 3. ROC points for malignant rules from ILP and decision tree classifier



ILP

BETTER THAN

DT

Conclusions

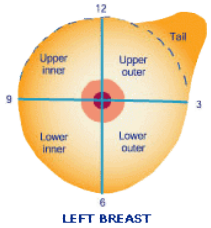
- We explored **alternatives** to our **best SVM** classifier and have shown that it is possible to **obtain** more **interpretable classifiers** with **same performance** on the test set

ILP EQUAL TO SVM

- We can generate **interpretable classifiers** with **higher performance** than our **best decision tree** classifier

ILP BETTER THAN DT

- We concluded that **Side**, **Clockface**, **Depth** and **Quadrant** are **not relevant** variables for **our dataset**



Future Work

- Search for **smoothing function** that can produce **less discrete results** for **ILP**

- Apply same techniques and methodology presented in this work to **larger** and **more varied datasets**

Keel Repository [5]

GEO Datasets [6]

TCGA Datasets [7]

Thanks

Questions?

Appendix

References

[1] N. B. C. Foundation. (2016) Breast Cancer Statistics. [Online]. Available: http://www.breastcancer.org/symptoms/understand_bc/statistics

[2] P. Ferreira, N. A. Fonseca, I. de Castro Dutra, R. W. Woods, and E. S. Burnside, "**Predicting malignancy from mammography findings and image-guided core biopsies**", IJDMB, vol. 11, no. 3, pp. 257–276, 2015. [Online]. Available: http://dx.doi.org/10.1504/IJDMB.2015.067319

[3] E. S. for Radiotherapy and Oncology. (2016) Handbook of brachytherapy. [Online]. Available: http://www.estro.org/binaries/content/assets/estro/about/gec-estro/handbook-of-brachytherapy/j-18-01082002-breast-print proc.pdf

[4] M. H. Amer, "**Genetic factors and breast cancer laterality**", Cancer Manag Res, vol. 16, no. 6, pp. 191–203, April 2014.

[5] Keel Dataset Repository. (2016). [Online]. Available: http://sci2s.ugr.es/keel/datasets.php

[6] GEO Datasets. (2016). [Online]. Available: https://www.ncbi.nlm.nih.gov/gds

[7] TCGA Datasets. (2016). The Cancer Genoma Atlas. [Online]. Available: https://cancergenome.nih.gov/