Comparing Analytical Decision Support Models Through Boolean Rule Extraction: A Case Study of Ovarian Tumour Malignancy

M.S.H. Aung[1], P.J.G Lisboa[1], T.A. Etchells[1], A.C. Testa[2], B. Van Calster[3], S. Van Huffel[3], L. Valentin[4], and D. Timmerman[5]

[1] School of Computing and Mathematical Sciences, Liverpool John Moores University, UK
{M.S.Aung,P.J.Lisboa,T.A.Etchells}@ljmu.ac.uk
[2] Istituto di Clinica Ostetrica e Ginecologica, Università Cattolica del Sacro Cuore, Rome, Italy
[3] Dept of Electical Engineering, ESAT-SCD, Katholieke Universiteit Leuven, Belgium
{ben.vancalster,sabine.vanhuffel}@esat.kuleuven.be
[4] Dept of Obstetrics and Gynaecology, University Hospital Malmö, Sweden
[5] Dept of Obstetrics and Gynaecology, University Hospitals, Katholieke Universiteit Leuven, Belgium
dirk.timmerman@uz.kuleuven.ac.be

Abstract. The relative performances of different classifiers applied to the same data are typically analyzed using the Receiver Operator Characteristic framework (ROC). This paper proposes a further analysis by explaining the operation of classifiers using low-order Boolean rules to fit the predicted response surfaces using the Orthogonal Search Based Rule Extraction algorithm (OSRE). Four classifiers of malignant or benign ovarian tumours are considered. The models analyzed are two Logistic Regression models and two Multi-Layer Perceptrons with Automatic Relevance Determination (MLP-ARD) each applied to a specific alternative covariate subset. While all models have comparable classification rates by Area Under ROC (AUC) the classification varies for individual cases and so do the resulting explanatory rules. Two sets of clinically plausible rules are obtained which account for over one half of the malignancy cases, with near-perfect specificity. These rules are simple, explicit and can be prospectively validated in future studies.

1 Introduction

The logic behind the behaviour of parametric classification models is often described with reference to *decision boundaries* within the data space of N dimensions where N is the number of input attributes or variables. However the often complex morphology of the boundaries makes it difficult to describe them explicitly. For this reason parametric classification models are often treated as black boxes and only evaluated as such often using the Area Under the Receiver Operator Characteristic curve (AUC). This study shows how deeper insight can be gained about the decision boundaries if they are approximated using axis parallel hyper-cubes in the data space. The axis parallel morphology corresponds to a Boolean rule specifying the limits to each variable in the

D. Liu et al. (Eds.): ISNN 2007, Part II, LNCS 4492, pp. 1177–1186, 2007.
© Springer-Verlag Berlin Heidelberg 2007

hyper-cube. Consequently these statements are explanations for how a model classifies a data point into a particular class. While two models may have similar classification performances the extracted rules will show whether the models are functionally similar or not. The rule extraction method used for this analysis is the Orthogonal Search Based Rule Extraction algorithm (OSRE) [5]. This is a fast and principled approach to approximate hyper-cubes to the decision boundaries of selected models for the purpose of yielding low order mutually overlapping Boolean explanatory rules. This method has been used in a variety of applications [2] [4] including the medical domain. The explicit rules generated by the OSRE framework are especially important in safety critical domains such as decision support for clinical medicine, as they enable direct validation against expert knowledge.

This study is focused on a prospectively acquired dataset for ovarian cancer kindly provided by the International Ovarian Tumour Analysis group (IOTA) as part of collaborative research funded by FP6 Network of Excellence: BIOPATTERN *(www.biopattern.org).* The International Ovarian Tumour Analysis group (IOTA) is a network of 9 clinical and academic centres from Belgium, Sweden, Italy, France and the United Kingdom that collected the data used in this study known as the IOTA Phase I dataset. The group applied strict data acquisition protocols [6] specifically for the purpose of parametric modeling. They consist of demographic variables, clinical signs and measurements derived from Doppler ultrasound scans for a cohort of 1066 patients diagnosed with having either a benign or a malignant ovarian tumour. There are over 40 explanatory variables measured for each patient. The database has been the object of several studies undertaken to produce classification models for diagnosing malignancy [1] [3] [6].

Four models that produced models with AUC of over 0.9 [3] [6] were selected for further investigation with rule extraction: The first two models labeled as 'M1' and 'M2' are logistic regression models and the latter two labeled as 'M3' and M4' are neural network classifiers both based on a Bayesian Multi Layer Perceptron method.

2 Orthogonal Search Based Rule Extraction

Orthogonal Search-based Rule Extraction (OSRE) is a framework for automatic rule extraction and pruning which efficiently extracts low order rules from the smooth decision surfaces created by analytical classifiers. The OSRE framework uses the data that constructed the smooth model to search, in orthogonal directions, for hyper-cubes containing the regions of the data space for which the model prediction is in-class (Fig. 1). The hyper-boxes that capture in-class data are converted into conjunctive rules expressed by the boundary values for each covariate. As the algorithm is applied to each data item that the model predicts to be in-class, there are as many rules produced as there are predicted in-class data, rule set R_n. This large number of rules is then automatically pruned according to predefined performance criteria intended typically to maximize the proportion of data explained or the positive predictive value (PPV) of the rule set [5].

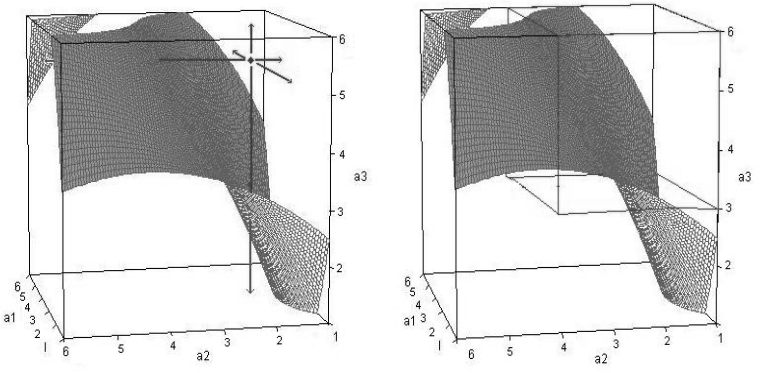

Fig. 1. An example of creating a hyper-cube: after orthogonally searching the data space from a sample point, the size of the cube is determined by the lengths of the search spans limited by the decision surface from the model or the extreme of the data space

All the rules are expressive in conjunctive form. For example the hyper-cube in Fig.1 can be stated as:

$$r_1 = (1 \le a1 \le 6) \wedge (1 \le a2 \le 4) \wedge (3 \le a3 \le 6), \quad (1)$$

where a_1, a_2 and $a3$ are input variables. The initial rule set is first reduced by removing any repeated rules. In some cases this single reduction strategy can reduce R_n to a very small rule set of conjunctive rules. However, if the data contains continuous variables then, in practice, further rule reduction techniques may be necessary.

The next phase of refinement is a minimum specificity filter. The rules whose specificity value is less than some predefined value are removed from the list. In clinical diagnostic applications we would only be interested in rules that have a very high specificity value, normally not accepting rules whose specificity is less than 0.9. After this process a minimum sensitivity filter is applied to give a measure of the *coverage* of the rule, i.e. how much of the in-class data the rule covers. Sensitivity and specificity values can now be calculated for the disjunction of all of the remaining rules. It is important to note that the rules can be mutually overlapping. The sensitivity and specificity values of the individual conjunctive rules or the combined set of disjunctive rules can be represented as steps in a Receiver Operator Characteristic (ROC) plot. The final goal of the rule refinement process is to achieve a rule set whose global ROC point is closest to the point with unit sensitivity and specificity.

3 Data and Classification Models

The *International Ovarian Tumour Analysis Group Phase I* dataset consists of 1066 patients with adnexal tumours. The data has been prospectively collected according to a standardized protocol at the nine IOTA group centres [6]. Much attention was given to uniformity of protocols to ensure integrity and consistency. The dataset consists of over 40 clinical, biochemical, and ultrasound variables that have the potential for malignancy diagnosis. The clinical variables included family history, age, and any previous hormonal therapy. Also included were variables derived from transvaginal

scans and transabdominal ultrasonography. The presence or absence of pain during the scan was also recorded. The target vector for the analytical models is the outcome of the tumour (benign or malignant).

3.1 Description of Multi-centre Models

Four classification models from previous multi-centre studies that have demonstrated good classification rates are selected for rule extraction. These were two logistic regression models described in Timmerman *et al* [6] and two Multi-Layer Perceptrons with Automatic Relevance Determination described in Van Calster *et al* [3]. Both studies aimed to produce models for the diagnosis of malignant ovarian tumours trained on the IOTA Phase I dataset. The variable selection methods and parameter determination algorithms are described in detail in these studies.

M1: Logistic Regression 1 with twelve variables selected
Timmerman *et al* [6] describes a linear logistic regression model with a 12 variable subset of the set of 40 variables (Table 1). Selection process is stepwise multivariate logistic regression. The AUC value of M1 for a validation set is 0.94 (SE =0.017).

M2: Logistic Regression 2 with a six variable subset
This model is a simplified version of M1 using six variables also presented in Timmerman et al [6], these variables being the first to be entered into the stepwise selection process of the M1 set. The AUC value of this for M2 is shown to be 0.92 (SE=0.018) for a validation set. The six variables are a subset of those in Table 1, namely Age, Ascites, PapFlow, MaxSolid, WallRegularity and Shadows.

M3: Multi-Layer Perceptron with Automatic Relevance Determination 1 with eleven variables selected
Described in Van Calster *et al* [3] this model uses the connectionist Multi Layer Perceptron neural network for prediction. The network utilises a Bayesian framework

Table 1. Variables used in Logistic Regression 1 (M1)

Variable Name	Description	Variable Type	Univariate *P* Value
PerHistOvCa	Personal History of Ovarian Cancer	Binary	0.0096
HormTherapy	Current Hormonal Therapy	Binary	0.0477
Age	Age of patient in years	Continuous	<0.0001
MaxLes	Size of largest lesion dimension in mm	Continuous	<0.0001
Pain	Presence of pelvic pain during scan	Binary	0.0034
Ascites	Presence of fluid outside Pouch of Douglas	Binary	<0.0001
PapFlow	Presence of blood flow within projections	Binary	<0.0001
Locul5	Morphologic category of lesion is solid	Binary	<0.0001
MaxSolid	Size of largest solid component dimension in mm	Continuous	<0.0001
WallRegularity	Whether internal wall of legion is irregular	Binary	<0.0001
Shadows	Presence of acoustic shadows	Binary	<0.0001
ColScore	Subjective assessment of amount of blood flow	4 Categories	<0.0001

that can automatically regularise the model and soft prune unwanted variables. This framework is known as Automatic Relevance Determination (ARD). Evaluation of the M3 model showed an AUC of 0.94 (SE=0.018) for a validation set.

Table 2. Variables used in MLP-ARD 1 (M3)

Variable Name	Description	Variable Type	Univariate *P* Value
PapNr	Number of Papillations	5 Categories	<0.0001
HormTherapy	Current Hormonal Therapy	Binary	0.0477
Age	Age of patient in years	Continuous	<0.0001
MaxLes	Size of largest lesion dimension in mm	Continuous	<0.0001
Locul4	Morphologic category of lesion is multi-locular solid	Binary	<0.0001
Ascites	Presence of fluid outside Pouch of Douglas	Binary	<0.0001
PapFlow	Presence of blood flow within projections	Binary	<0.0001
Locul5	Morphologic category of lesion is solid	Binary	<0.0001
MaxSolid	Size of largest solid component dimension in mm	Continuous	<0.0001
WallRegularity	Whether internal wall of legion is irregular	Binary	<0.0001
ColScore	Subjective assessment of blood flow	4 Categories	<0.0001

M4: Multi-Layer Perceptron with Automatic Relevance Determination 2 with eleven variables selected

Also in Van Calster *et al* [3] a second MLP-ARD model with a good classification rate is described with a differing variable set, with an AUC of 0.93 (SE=0.017) for a validation set, This model replaces the variables 'hormonal therapy', 'colour score', 'number of papillations' and 'locularity being: multi-locular solid' from M3 with 'pelvic pain', 'acoustic shadows', 'personal history of ovarian cancer' and 'locularity being unilocular'. Therefore, the variables for M4 are:

Table 3. Variables used in MLP-ARD 2 (M4)

Variable Name	Description	Variable Type	Univariate *P* Value
PerHistOvCa	Personal History of Ovarian Cancer	Binary	0.0096
Pain	Presence of pelvic pain during scan	Binary	0.0034
Age	Age of patient in years	Continuous	<0.0001
MaxLes	Size of largest lesion dimension in mm	Continuous	<0.0001
Locul1	Morphologic category of lesion is unilocular	Binary	<0.0001
Ascites	Presence of fluid outside Pouch of Douglas	Binary	<0.0001
PapFlow	Presence of blood flow within projections	Binary	<0.0001
Locul5	Morphologic category of lesion is solid	Binary	<0.0001
MaxSolid	Size of largest solid component dimension in mm	Continuous	<0.0001
WallRegularity	Whether internal wall of legion is irregular	Binary	<0.0001
Shadows	Presence of acoustic shadows	Binary	<0.0001

4 Results

Two disjunctive sets of rules were generated for each model with different rule refinement criteria. The PPV criterion limits misclassifications but the sensitivity

Table 4. Set 1 and Set 2 rules for malignancy in descending order of PPV

Rule Number	Conjunctive Rule Statements	For Individual Rule			Cumulative Value in Set		
		Sensitivity	Specificity	PPV	Sensitivity	Specificity	PPV
Set 1							
1	69 <= MaxLes <= 410 Pain = 0 44 <= MaxSolid <= 230 ColScore =4	0.18	1	1	0.18	1	1
2	54 <= Age <= 94 Pain = 0 Ascites = 1 WallRegularity = 1 Shadows = 0 ColScore =2 or 4	0.18	0.99	0.97	0.3	0.99	0.98
3	108 <= MaxLes <= 410 Pain = 0 42 <= MaxSolid <= 230 WallRegularity = 1	0.22	0.99	0.97	0.39	0.99	0.97
4	HormTherapy = 0 46 <= Age <= 94 PapFlow = 1 15 <= MaxSolid <= 230 ColScore = 3 4	0.19	0.99	0.95	0.49	0.99	0.96
5	31 <= Age <= 94 77 <= MaxLes <= 410 Pain = 0 30 <= MaxSolid <= 230 WallRegularity = 1 Shadows = 0 ColScore = 2 or 4	0.32	0.99	0.94	0.54	0.99	0.95
Set 2							
1	WallRegularity = 1 ColScore = 4 51.6<= Age <= 93.5	0.22	1	1	0.22	1	1
2	HormTherapy = 0 WallRegularity = 1 ColScore = 4 108<=MaxLes <=403	0.14	1	1	0.28	1	1
3	PapNr = 5 PapFlow = 1 ColScore = 3 4 59<=MaxLes <= 401 19.9<=MaxSolid<=227	0.18	1	1	0.39	1	1
4	WallRegularity = 1 Locul5 = 1 49.72<= Age <= 92.25	0.20	0.99	0.96	0.49	0.99	0.98
5	HormTherapy = 0 PapNr = 3 4 5 ColScore = 3 4 28.7<=MaxSolid<=224.2	0.17	0.99	0.97	0.52	0.99	0.97

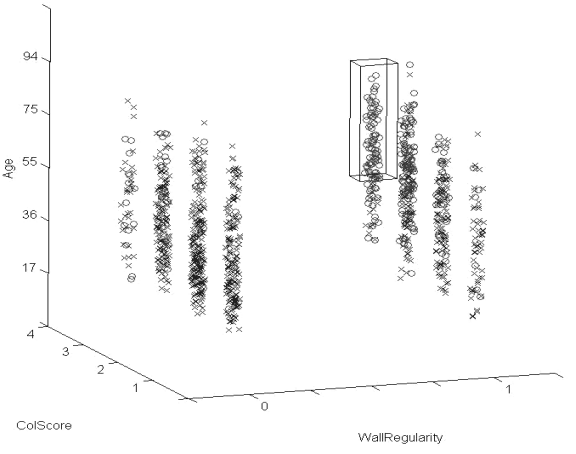

Fig. 2. Visualization of the hyper-cube for Rule 1 from Set 2 which only contains malignant cases represented by circles

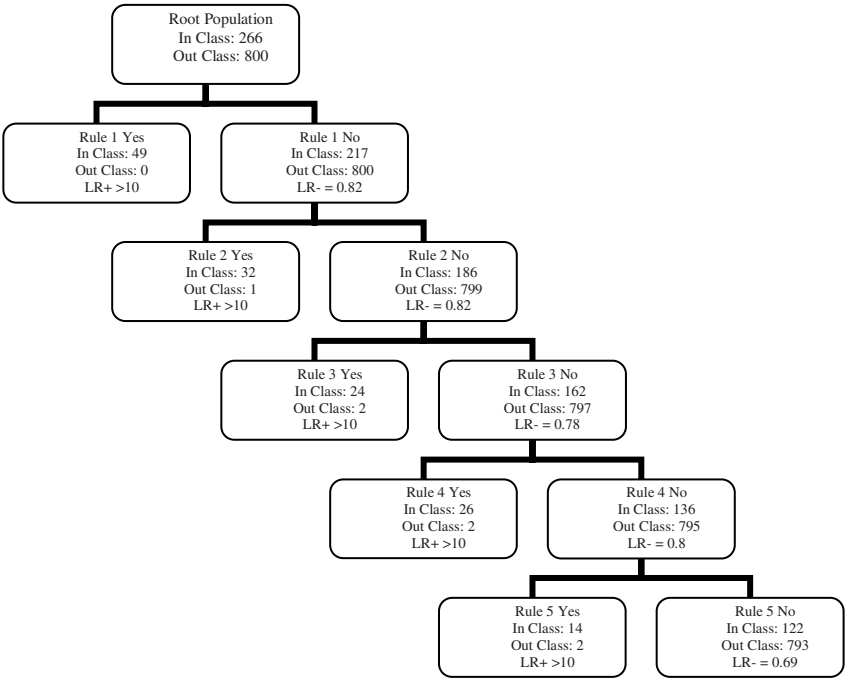

Fig. 3. Hierarchical Rule Tree for Set 1: The Positive and Negative Likelihood Ratios for each rule is shown

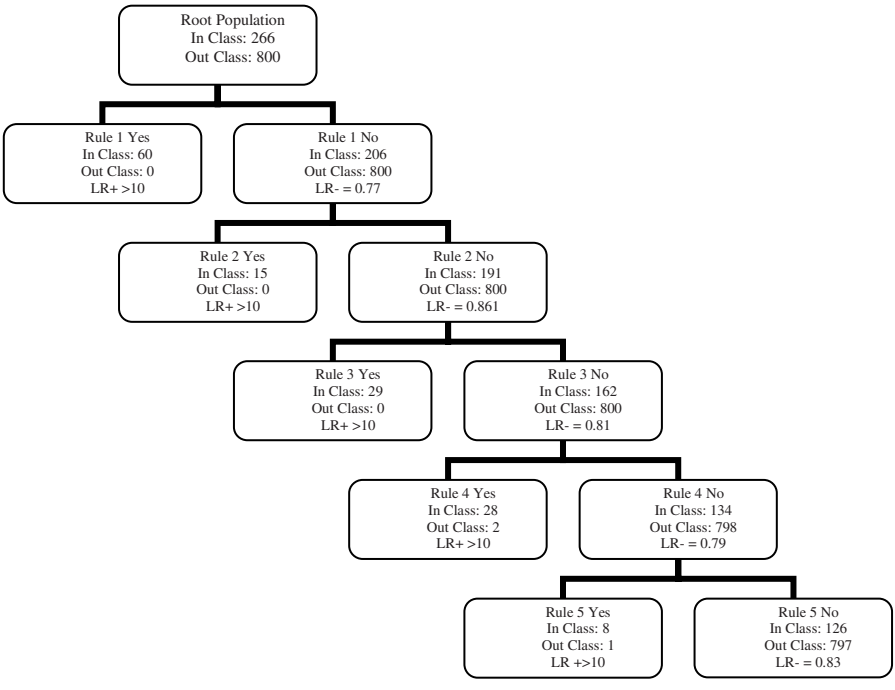

Fig. 4. Hierarchical Rule Tree for Set 2: The Positive and Negative Likelihood Ratios for each rule is shown

criterion leads to better coverage of the data. The PPV criterion resulted in better specificity for the overall disjunctive rule set, therefore this was the preferred approach. Moreover two of the analytical models (M1 and M3) generated rules with near perfect specificity. The ranked rules were then truncated when too few additional cases were explained. This resulted in the two rule sets listed in Table 4. Hierarchical rule trees describing the rule sets and their performance in mutually exclusive branches are shown in Fig. 3 and Fig. 4.

All rules from both rule sets showed high specificity and thus a high positive likelihood (LR+) ratio for the detection of malignancy are given by:

$$LR+ = \frac{Sensitivity}{1 - Specificity} \tag{2}$$

and the negative likelihood ratio (LR-) is also calculated to quantify the predictive power of the rule complement for detecting benign cases, given by:

$$LR- = \frac{1 - Sensitivity}{Specificity} \cdot \tag{3}$$

Rule Set 1 cumulatively identifies 144 of the 266 malignant cases with 2 misclassifications, compared to set 2 which identifies 140 malignant cases with 3 misclassifications. The overall sensitivities of Set 1 and 2 are 0.54 and 0.52, respectively. From the

ten selected rules shown there are 4 rules that did not yield any misclassifications and hence have a specificity of 1, one of these is Rule 1 from Set 2 visualized in Fig. 2. The common variable in these four rules is the subjective categorization of the amount of blood flow 'ColScore'. This demonstrates the value of this variable despite the inherent uncertainty due to its subjective nature. All of the rules were assessed by a clinical expert and found to be consistent with current knowledge.

5 Conclusions

The operation of alternative high performing classifiers was described using multivariate hierarchical rule trees with few low order rules. This analysis framework is proposed as the basis for direct validation of analytical models against domain expertise thus complementing and extending the traditionally used quantitative indices of aggregate performance.

Over half of the malignant cases are explained by each rule set. The remaining cases may require a different classification paradigm whereby observations are allocated into ranked groups for risk of malignancy, so to describe the clinical profiles representing more difficult classifications, which fall at intermediate values of the risk of malignancy. This approach is currently being developed.

6 Discussion

Rules are potentially more useful to the clinician than mathematical models because they can be taught and learned and be used without computation. On the other hand they might prove useful to discover previously undetected clinical correlations and they might give more insight in the ultrasound morphology and clinical presentation of benign and malignant ovarian tumours. While simple rules based on single variables seem not to be reliable or seem to be useful only to a small proportion of the patients, systematic rule extraction offers a much more extensive exploration of all possible rules hidden in a database. However, their clinical value needs to be assessed in future studies. A prospective validation is planned in IOTA Phase II.

References

1. Ameye, L: Predictive Models for Classification in Gynecology. PhD thesis, Faculty of Engineering, Katholieke Universiteit Leuven, Leuven, Belgium (2005)
2. Aung, M.S.H., Lisboa, P.J.G., Taktak, A.F.G., Damato, B.E: Modelling Survival of Intraocular Melanoma using a Partial Logistic Artificial Neural Network with Automatic Relevance Determination and Orthogonal Search Based Rule Extraction. Proc. Computational Intelligence in Medicine (CIMED), Lisbon, Portugal (2005) 114-121
3. Van Calster, B., Timmerman, D., Nabney I.T., Valentin, L., Van Holsbeke, C., Van Huffel, S: Classifying Ovarian Tumors using Bayesian Multi-Layer Perceptrons and Automatic Relevance Determination: A Multi-Center Study, in Proc. of the 28th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC2006), New York City, USA (2006) 5342-5345

4. Etchells, T.A., Harrison, M.J: Orthogonal Search-Based Rule Extraction for Modelling the Decision to Transfuse. Anesthesia **61** (4) (2006) 335-338
5. Etchells, T.A., Lisboa, P.J.G: Orthogonal Search-Based Rule Extraction (OSRE) for Trained Neural Networks: A Practical and Efficient Approach. IEEE Trans Neur. Net. **17** (2) (2006) 734-384
6. Timmerman, D., Testa, A.C., Bourne, T., Ferrazzi, E., Ameye, L., Konstantinovic, M.L., Van Calster, B., Collins, W.P., Vergote, I., Van Huffel, S., Valentin, L: Logistic Regression Model to Distinguish Between the Benign and Malignant Adnexal Mass Before Surgery: A Multicenter Study by the International Ovarian Tumor Analysis Group. J Clin Oncol. (2005) 8794-8801