# Preoperative prediction of malignancy of ovarian tumours using modified sequential non-uniform procedure

Viktoriya Stalbovskaya, Emmanuel C. Ifeachor, Sabine Van Huffel, and Dirk Timmerman

*Abstract*— In this paper, we present an extension of sequential non-uniform procedure (SNuP) with application of the method to ovarian tumour data, obtained during multicentre study by the International Ovarian Tumour Analysis Group (IOTA). The inference method combines feature selection based on the Kullback information gain and a step-wise classification procedure to produce a reliable, interpretable and robust model. In particular, we extend SNuP to enable it to handle continuous variables without the need for manual specification of thresholds. We applied the extended model to an ovarian tumour data set to distinguish between malignant and benign tumours. The performance of the model was assessed using ROC analysis and gave 86.9% of sensitivity and 84.3% of specificity with overall accuracy level of 84.9%.

#### I. INTRODUCTION

Ovarian tumours are common among women. In Europe and North America the age-adjusted standardised incidence rate of ovarian cancer is over 10 per 100,000 women [1]. Preoperative prediction of malignancy of ovarian tumours is very important, because it can prevent unnecessary surgery for benign functional cysts or in the case of benign neoplastic lesions only minimal surgical intervention would be required. On the other hand, patients with malignant forms of tumour require not only surgical operation but also an appropriate pre-, peri- and postoperative management. A great deal of effort has been put in by gynecological oncologists in order to develop preoperative predictive markers of ovarian malignancy. However, prospective testing of these markers have shown either low performance or unbalanced results (i.e. high specificity and low sensitivity). To address the limitations of previous studies the International Ovarian Tumour Analysis (IOTA) Group have established multicentre prospective clinical trial with more than six centres working to the same protocol [2] and collecting data from a total of 1000 patients who have a persistent adnexal mass.

For clinical acceptance, a predictive model for discrimination of ovarian tumours should satisfy the following requirements:

- (i) have reasonably high sensitivity and specificity levels, typically 90% and 75%, respectively [3],
- (ii) be interpretable, and
- (iii) use as few diagnostic techniques/parameters as possible.

In relation to (iii), the range of laboratory and instrumental diagnostic techniques for ovarian cancer is wide and includes transvaginal and transabdominal ultrasonography, serum tumour markers, laparoscopy, computer tomography and magnetic resonance imaging. A key problem is in the choice of necessary procedures taking into account their diagnostic value, cost and invasiveness.

In [4] we described a method of modelling the preoperative diagnosis of ovarian tumours based on the Sequential Nonuniform Procedure (SNuP), which meets the requirements above. SNuP is based on the Naïve Bayes classification, but with additional restrictions. In particular, consecutive multiplication of likelihood ratios of input variables is interrupted when one of the diagnostic thresholds [5] is reached. Values of thresholds are specified according to an acceptable level of the diagnostic errors.

The SNuP operates sequentially on the variables (features) as the cases (observations) are accumulated. This is important because it makes it possible to personalise differential diagnosis. This is achieved by varying the number of attributes used, ranking the variables according to their discriminative relevance and the specified confidence level.

The current SNuP does not handle continuous variables and requires manual specification of thresholds for this type of input features. This paper presents an approach to overcome this problem by univariate and multivariate clustering of continuous variables and the application of the modified SNuP to IOTA phase I data set (see later).

The remainder of the paper is organised as follows. In Section II, the data and method are presented. In Section III, results of analysis are given followed by a performance evaluation. Finally, we present the conclusion and future work.

# II. METHODS AND PATIENTS

# A. Methods of analysis

The key issue in preoperative diagnosis is to determine whether a given patient belongs to one of two groups: benign or malignant tumour, given the symptoms and laboratory data. The task can be viewed as a two-class classification ( $A_k$ , where k = 1, 2) problem, given a vector of input variables, **x**.

Let us denote  $P(A_k)$  - prior probability of class k, k = 1...n and n is number of classes (groups),  $P(x_i|A_k)$  - conditional probability of  $x_i$  given  $A_k$ , i.e. probability of presence of symptom  $x_i$  in the group  $A_k$ ,  $P(x_i)$  - prior probability of symptom  $x_i$ . So the posterior probability of the patient to belong to group  $A_k$  having symptom  $x_i$  can be defined using Bayes' theorem:

$$P(A_k|x_i) = \frac{P(A_k)P(x_i|A_k)}{\sum_k P(A_k)P(x_i|A_k)}$$
(1)

Sequential non-uniform procedure produces a model of classification into two groups. The ratio of conditional probabilities of the groups is equal to the ratio of symptom's occurrences in the two groups.

$$\frac{P(A_1|x_i)}{P(A_2|x_i)} = \frac{P(x_i|A_1)}{P(x_i|A_2)}$$
(2)

where  $\frac{P(A_1|x_i)}{P(A_2|x_i)}$  is a likelihood ratio of probability of a group given symptom  $x_i$ ,  $\frac{P(x_i|A_1)}{P(x_i|A_2)}$  is a likelihood ratio of the probability of the symptom  $x_i$  given groups  $A_k$ .

Accumulation of the diagnostic information given the presence of independent features/symptoms  $x_1, x_2, ..., x_n$  is performed as

$$\frac{P(A_1|x_1, x_2, \dots, x_n)}{P(A_2|x_1, x_2, \dots, x_n)} = \prod_{i=1}^n \frac{P(x_i|A_1)}{P(x_i|A_2)}$$
(3)

The *inference process* uses two types of errors to determine the thresholds for the ratio in (3) to make a decision -  $\alpha$  and  $\beta$ . In terms of 'malignant-benign' classification,  $\alpha$  specifies the probability of false assignment of a patient with malignant tumour into a benign tumour group, and  $\beta$  specifies the probability of false assignment of a patient with benign tumour to a malignant tumour group. In terms of classification into groups  $A_1$  and  $A_2$ ,  $\alpha$  is the rate of misclassification in group  $A_1$ ,  $\beta$  is the rate of misclassification in group  $A_2$ .

A diagnostic coefficient  $DC_i$  of symptom  $x_i$  is a score value which is defined as

$$DC_i = 10 \log_{10} \frac{P(x_i|A_1)}{P(x_i|A_2)}$$
(4)

Accumulation of the diagnostic information using the diagnostic coefficients is performed as a sum:

$$\sum DC(x_i) = DC(x_1) + DC(x_2) + \dots + DC(x_n)$$
 (5)

The threshold for a diagnostic hypothesis is the minimum acceptable rate of correct diagnoses over incorrect ones. Thresholds for the sum of the diagnostic coefficients are defined as

$$DC_{th}(A_1) = 10\log_{10}\frac{1-\alpha}{\beta}$$
 (6)

$$DC_{th}(A_2) = 10 \log_{10} \frac{\alpha}{1-\beta}$$
 (7)

where  $\alpha$  and  $\beta$  are acceptable levels of diagnostic errors for group  $A_1$  and  $A_2$  respectively.

The SNuP using diagnostic coefficients is performed until the following inequality is true. When the inequality is broken the diagnostic decision is made.

$$DC_{th}(A_2) < \sum_i DC(x_i) < DC_{th}(A_1)$$
(8)

The feature selection process and ranking of input variables/symptoms is based on the calculation of symmetrised Kullback-Leibler divergence between two distributions, P and Q, so-called J-divergence [6]. The J-divergence of the *distinct* value of the variable is defined as

$$J(x_{ij}) = \frac{P(x_{ij}|A_1) - P(x_{ij}|A_2)}{2} 10 \log_{10} \frac{P(x_{ij}|A_1)}{P(x_{ij}|A_2)}$$
(9)

The *J*-divergence of the *variable* is the sum of the information measures of all its distinct values:

$$J(x_i) = \sum_{j=1}^{m} J(x_{ij})$$
(10)

In order to obtain the most informative features the information measures were calculated for all variables and then sorted in a descending order according to the *J*-value.

*Continuous variables* were transformed to ordinal by partitioning the initial input space. Thereafter these variables were analysed by SNuP in the regular way. Automatic partition of the input space for continuous variables was performed by applying k-means clustering [7] with three number of clusters. We used squared Euclidean distance as a distance measure, initial centroid positions of clusters were selected randomly. In case of a cluster losing all of its member observations those clusters were removed. Assignment of continuous variables from the test set to clusters was made on a basis of minimal squared Euclidean distance to one of the centroids that were identified on the training stage.

An algorithm for building the decision rule for differential diagnosis involves the following steps:

- 1) calculate values of diagnostic coefficients  $DC_i$  for all symptoms  $x_i$  using (4).
- 2) calculate *J*-divergence for all  $x_i$  using (10) and sort values in a descending order.
- specify an acceptable level of errors, α and β, and calculate the thresholds for diagnostic coefficients from (6) and (7).
- start the process of accumulation of diagnostic information according to (5).
- 5) end the process when the inequality (8) is broken or there are no variables left.

The performance of the model was assessed by calculating the parameters of receiver operating characteristics: overall accuracy (Acc), sensitivity (Se), specificity (Sp), predictive positive value (PPV), predictive negative value (PNV). Evaluation of the model was performed by applying a 3-fold cross validation. The initial data set was split randomly into a training set and a test set with the proportion of malignant to benign cases equal to 1:3. The results were summarised as mean and standard deviation for Acc, Se, Sp, PPV, and PNV.

# B. Patients

The IOTA I data set was collected in multicentre prospective clinical study with a common protocol [2]. Full database include 1066 cases of ovarian tumours, 266 malignant and 800 benign. Histological diagnosis were used as a gold standard. There were three data modalities: (i) clinical variables included family history of ovarian and breast cancer, age, menopausal status, previous hormonal surgery, and surgical history; (ii) sonographical examination was performed in all cases with gray scale and colour Doppler imaging with total over 40 morphological and blood flow velocity characteristics; (iii) serum tumour marker CA125 was measured for 809 patients. When intratumoral blood flow velocity waveforms were not detected, the peak systolic velocity (PSV), time averaged maximum velocity (TAMXV), the pulsatility index (PI), and the resistance index (RI) were substituted by 2.0 cm/sec, 1cm/sec, 3.0, and 1.0, respectively [3]. Descriptive statistics and univariate analysis of ultrasound characteristics are given in a previous publications on analysis of IOTA phase I study [3], [8].

# **III. RESULTS**

#### A. Transformation of data

At the first stage of analysis some preprocessing procedures were made in order to incorporate continuous variables into a model. Continuous variables were transformed into discrete by automatical partitioning input space into intervals. We used univariate and multivariate k-means clustering procedure in MATLAB for that. For volumetric characteristics such as diameter of the lesion (LesD1-3) was used multivariate clustering. Table below demonstrates this approach. Firstly, one has to specify variables and desired number of clusters. We used three number of clusters by default, considering that values in these clusters might be described as low, medium and high. As a result, one obtain new variable with three/two values. Assignment of values for the variable was made by this rule

$$value = \arg_i \min \sum_{j=1}^{m} (x_j - c_{i,j})^2$$
 (11)

where  $\sum_{j=1}^{m} (x_j - \mathbf{c}_{i,j})^2$  is Euclidean distance from an observation x to one of m-dimensional centroids c of clusters. So in the case of malignant-benign classification using diameter of the lesion number of clusters  $\max(i) = 3$ , number of modalities  $\max(j) = m = 3$  and coordinates of clusters' centroids  $\mathbf{c}_i$  are triplets {LesD1; LesD2; LesD3}<sub>i</sub>: {51.0; 40.7; 40.0}, {106.8; 85.3; 82.2}, {201.2; 148.0; 142.3}.

Variable	Values					
variable	Cluster 1	Cluster 2	Cluster 3			
LesD1	51.0	106.8	201.2			
LesD2	40.7	85.3	148.0			
LesD3	40.0	82.2	142.3			
LesD	1	2	3			
lg(CA125)	1.11	1.83	3.01			
lg(CA125)	1	2	3			

# B. Classification of ovarian tumours using SNuP

Please refer to [4] for a few examples of application of the method to differential diagnosis of benign and malignant forms of ovarian tumours.

After all input variables are measured in discrete scale the elements of SNuP model are calculated. In comparison with

our previous paper [4], where only binary variables were used, here we used the whole range of unique values of ordinal and nominal variables. For every distinct value of a variable (e.g. strong blood flow,  $col_score = 4$ )there were calculated the following parameters: conditional probability of this event in malignant and benign groups  $P(x_i|A_{1,2})$ , diagnostic coefficient for the distinct value of the variable  $DC(x_{i,j})$  using (4), J-divergence of the symptom's level  $J(x_{i,j})$  using (9). Then values of J-divergence were summarised across all values to produce total estimate of informativity of the symptom  $J(x_i)$ as in (10). It is recommended to start SNuP from the most informative variables, therefore all variables were sorted by  $J(x_i)$  value in descending order. As the format of the paper does not allow to present DC and J for all symptoms, we summarised the most informative variables in Table I where total J is presented in the third column, all distinct values of variables are given in the forth one, followed by corresponding values of diagnostic coefficients. A large value of DC means a high discriminative ability of the variable and the information measures J gives an indication of how reliable this is. Features with positive DC values correspond to malignancy, and those with negative values to the benign group. Accumulation of the diagnostic information was carried out by summation of the diagnostic coefficients and comparing the sum with a specified threshold.

In order to classify cases there are thresholds  $DC_{th}(A_{1,2})$  to be specified. We set  $\beta$  to 0.05 and varied  $\alpha$  from 0.90 to 0.001. Lower and upper thresholds for sum of diagnostic coefficients were calculated using (6) and (7). Performance of the model during 3-fold cross-validation is presented in Table II. Last column of the table shows median number of cases where diagnostic decision was undefined. As can be seen this number increases with decreasing the level of acceptable error  $\alpha$ . We chose  $\alpha = 0.10$  as an optimal threshold as it produce relatively high performance (Se=86.9%, Sp=84.3%, Acc=84.9%) and low number of undefined cases (10 out of 355).

Interpretation of diagnostic coefficients of untransformed and univariately transformed variables is straightforward when there is clear assignment of DC to the level of the symptom. For instance, low blood flow (ColScore=1) is highly associated with benign tumour, DC = -11.2, and strong blood flow (ColScore=4) on the other hand is a marker of malignancy, DC = 9.6. Example of univariately transformed variable, log of serum CA125, was split into three clusters which can be described as low, medium and high level with raising degree of association with malignancy.

New variables created in two- or multidimensional space might have increasing values in one dimension and decreasing in another which slightly complicates interpretation and requires more clinical input for that. There are examples of these kind of variables: diameter of solid component (SolidD), velocity indices (PI, RI, PSV, TAMXV), diameter of ovaries (OvD) and diameter of lesion (LesD).

Rank	Variable	J	Level of symptom	Diagnostic coefficients
1	Locularity	5.37	{1 2 3 4 5 6}	$\{-15.9; 2.4; -6.1; 3.8; 7.4; 4.7\}$
2	ColScore	3.91	$\{1 2 3 4\}$	$\{-11.2; -3.6; 2.0; 9.6\}$
3	WallRegularity	2.76	{0 1}	$\{-6.4; 4.2\}$
4	SolidD	2.68	{1 2}	$\{-2.6; 9.9\}$
5	Ascites	2.64	{0 1}	$\{-2.1; 11.9\}$
6	PI, RI, PSV, TAMXV	2.19	{1 2 3}	$\{11.6; 5.2; -3.3\}$
7	RatioPapLes	2.05	$\{1\ 2\ 3\}$	$\{6.5; -2.8; 7.5\}$
8	NrLocules	1.99	$\{0\ 1\ 2\ 3\ 4\ 5\ 6\}$	$\{6.6; -4.0; -0.3; -2.9; -1.5; 1.8; 7.7\}$
9	Fluid	1.79	{1 2}	$\{8.9; -1.9\}$
10	PapNr	1.78	$\{0\ 1\ 2\ 3\ 4\}$	$\{-1.9; -0.7; 5.3; 2.7; 11.1\}$
11	PapFlow	1.68	$\{0\ 1\}$	$\{-1.9; 8.1\}$
12	MaxSolid	1.63	{1 2 3}	$\{-1.4; 9.1; 11.5\}$
13	age	1.35	{1 2 3}	$\{-4.4; 4.8; 0.9\}$
14	OvD	1.3	$\{1\ 2\ 3\}$	$\{-3.2; 6.3; 1.8\}$
15	PapSmooth	1.21	{0 1}	$\{-1.7; 6.4\}$
16	MaxLes	1.19	{1 2 3}	$\{-3.0; 6.1; 2.1\}$
17	lg(CA125)	1.11	$\{1 \ 2 \ 3\}$	$\{0.7; 4.8; 13.6\}$
18	LesD	1.05	$\{1 \ 2 \ 3\}$	$\{-2.8; 2.7; 5.3\}$

TABLE I The most informative variables with J>1.0

TABLE II Performance of the model on the test data set (  $\beta=0.05)$ 

α	Acc, %	Se, %	Sp, %	PPV, %	NPV, %	# undefined
0.90	68.7	91.7	61.2	44.4	95.5	0
0.80	69.1	90.2	62.2	45.0	94.9	0
0.70	68.4	90.5	61.0	44.6	94.9	0
0.60	68.5	91.3	61.0	44.8	95.2	0
0.50	73.4	87.9	68.5	50.4	94.3	0
0.40	76.7	87.5	73.2	53.9	94.4	0
0.30	79.4	87.4	76.8	56.5	94.8	3
0.20	81.4	86.5	79.7	59.2	94.7	6
0.10	84.9	86.9	84.3	65.3	95.2	10
0.05	86.5	86.6	86.4	67.7	95.3	13
0.01	90.3	86.2	91.6	76.6	95.5	32
0.001	92.8	84.5	95.2	84.3	95.5	50

# IV. CONCLUSION AND FUTURE WORK

In the paper we have presented the sequential non-uniform procedure for modelling the preoperative diagnosis of adnexal masses collected during multicentre prospective clinical trial by IOTA group. The method can be considered as an extension of Wald's consecutive analysis where accumulation of diagnostic information makes it possible to use a minimal number of features in order to make a decision with any given level of confidence. The advantages of the method include the use of a minimal number of variables, permissibility of cases with missing values, and interpretability of the model and results. The model can also incorporate prior knowledge of the distribution of the classes. From the end-user point of view (i.e. clinician) SNuP produces a model which is understandable and enables ranking of the input variables according to their discriminative abilities.

In this paper we extend SNuP to enable it to handle continuous variables without the need for manual specification of thresholds. It was done by using k-means clustering for automatic partitioning of the input space. A subject of the future work will be on multiclass generalisation [9], [10] and less strict handling of continuous variables.

#### ACKNOWLEDGMENTS

We acknowledge the financial support of the European Commission (The BIOPATTERN Project, Contract No. 508803) for this work.

#### REFERENCES

- [1] "Cancer incidence in five continents. Volume VIII," *IARC Sci Publ*, no. 155, pp. 1–781, 2002.
- [2] D. Timmerman, L. Valentin, T. Bourne, W. Collins, H. Verrelst, I. Vergote, and International Ovarian Tumor Analysis (IOTA) Group, "Terms, definitions and measurements to describe the sonographic features of adnexal tumors: a consensus opinion from the international ovarian tumor analysis (iota) group," *Ultrasound Obstet Gynecol*, vol. 16, no. 5, pp. 500–5, Oct 2000.
- [3] D. Timmerman, A. C. Testa, T. Bourne, E. Ferrazzi, L. Ameye, M. L. Konstantinovic, B. Van Calster, W. P. Collins, I. Vergote, S. Van Huffel, and L. Valentin, "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*, vol. 23, no. 34, pp. 8794–8801, Dec 2005.
- [4] V. Stalbovskaya, E. C. Ifeachor, S. Van Huffel, and D. Timmerman, "Modelling the preoperative diagnosis of ovarian tumours using sequential non-uniform procedure," *IEEE Transactions on Biomedical Engineering, accepted*, 2006.
- [5] E. Gubler, Computational methods of analysis and recognition of pathological processes. Medicina, Moscow, 1978.
- [6] H. Jeffreys, "An invariant form of the prior probability in estimation problems," J. R. Statist. Soc., vol. A, pp. 453–469, 1946.
- [7] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, 1967, pp. 281–297.
- [8] L. Valentin, L. Ameye, A. Testa, F. Lecuru, J. Bernard, D. Paladini, S. Van Huffel, and D. Timmerman, "Ultrasound characteristics of different types of adnexal malignancies," *Gynecol Oncol*, vol. 102, no. 1, pp. 41–8, Jul 2006.
- [9] C. Baum and V. Veeravalli, "A sequential procedure for multihypothesis testing," *IEEE Trans Inform Theory*, vol. 40, no. 6, pp. 1994–2007, Nov 1994.
- [10] V. P. Dragalin, A. G. Tartakovsky, and V. V. Veeravalli, "Multihypothesis sequential probability ratio tests - part i: Asymptotic optimality." *IEEE Trans Inform Theory*, vol. 45, no. 7, pp. 2448–2461, 1999.