Group Penalized Logistic Regressions Predict Ovarian Cancer

Ying Xie (wdmzjxieying@163.com)
https://orcid.org/0000-0002-5131-5526

Research Article

Keywords: Ovarian benign tumors and malignant tumors, Group coordinate descent algorithm, Group LASSO/SCAD/MCP estimator

Posted Date: January 3rd, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1223870/v1

Group Penalized Logistic Regressions Predict Ovarian Cancer *

Xuemei Hu[a], Ying Xie[a*], Yanlin Yang[b], Huifeng Jiang[b*]
a. School of Mathematics and Statistics,
b. Research Center for Economy of Upper Reaches of the Yangtse River,
Chongqing Technology and Business University,Chongqing 400067,China

Objectives: Ovarian cancer ranks first among gynecological cancers in terms of the mortality rate. Accurately diagnosing ovarian benign tumors and malignant tumors is of immense important. The goal of this paper is to combine group LASSO/SCAD/MCP penalized logistic regression with machine learning procedure to further improve the prediction accuracy to ovarian benign tumors and malignant tumors prediction problem.

Methods: We combine group LASSO/SCAD/MCP penalty with logistic regression, and propose group LASSO/SCAD/MCP penalized logistic regression to predict the benign and malignant ovarian cancer. Firstly, we select 349 ovarian cancer patients data and divide them into two sets: one is the training set for learning, and the other is the testing set for checking, and then choose 46 explanatory variables and divide them into 11 different groups. Secondly, we apply the training set and group coordinate descent algorithm to obtain group LASSO/SCAD/MCP estimator, and apply the testing set to compute confusion matrix, accuracy, sensitivity and specificity. Finally, we compare the prediction performance for group LASSO/SCAD/MCP penalized logistic regression with that for artificial neural network (ANN) and support vector machine (SVM).

Results: Group LASSO/SCAD/MCP/ penalized logistic regression selects 6/4/1 groups. The prediction accuracy and AUC for group MCP/SCAD/LASSO penalized logistic regression/SVM/ANN is 93.33%/85.71%/82.26%/74.29%/72.38% and 0.892/0.852/0.823/0.639/0.789, respectively.

Conclusions: Group MCP/SCAD/LASSO penalized logistic regression performs than SVM and ANN in terms of prediction accuracy and AUC. In particular, group MCP penalized logistic regression predicts the best. Therefore, we suggest group MCP penalized logistic regression to predict ovarian tumors.

Keywords: Ovarian benign tumors and malignant tumors; Group coordinate descent algorithm; Group LASSO/SCAD/MCP estimator

*Corresponding author：Ying Xie, E-mail:wdmzjxieying@163.com. Xuemei Hu, Professor in Statistics, School of Mathematics and Statistics, Chongqing Technology and Business University; PhD in Probability and Mathematical Statistics, School of Mathematics and Statistics, Central South University; Post-doctorate in System Science, Academy of Mathematics and Systems Science, Chinese Academy of Sciences; Visiting scholar, London School of Economics and Political Science.

1 Introduction

Ovarian cancer is a malignant tumor growing on the ovary. Its incidence rate is lower than that of cervical and endometrial cancer, whereas its mortality rate is higher than the sum of both cervical cancer and endometrial cancer and ranks first among gynecologic cancers. According to the global cancer data released by the World Health Organization International Agency in 2020, there were 314,000 new cases and 207,000 deaths of ovarian cancer in the world including 55,000 new cases and 38,000 deaths in China. Ovarian tumors are usually not difficult to diagnose. But benign and malignant diagnosis is not easy. Correct benign and malignant diagnosis need further auxiliary examinations, such as ultrasonography, cytology, laparoscopy, determination of tumor markers, radiologic diagnosis, etc.. Computer tomography (CT) and magnetic resonance imaging (MRI) can clearly show the image of the tumor, and play a vital role in the diagnosis of ovarian tumors, in the observation of residual neoplastic changes at any time, and in the recurrence of the tumor. Positron emission tomography (PET) is also helpful in differentiating benign and malignant tumors and in diagnosing recurrent cancers. PET-CT has the dual functions of both CT and PET, and help better diagnosis. Advanced ovarian cancer is the leading cause of cancer death in women. Especially after chemotherapy, the recurrence rate was still over 70%. High degree of malignancy, high recurrence rate and poor prognosis from advanced ovarian cancer has become some prominent factors affecting the survival of ovarian cancer patients. Therefore, it is crucial to accurately diagnose ovarian benign tumors and malignant tumors.

There are many studies related to ovarian cancer, including the influencing factors analysis, screening methods, tumor markers, treatment and prognosis. Kikkawa et al. (1998) assessed the value of tumor markers and clinical characteristics in making a differential diagnosis between MCT and squamous cell carcinoma arising from MCT, demonstrated that there were significant differences in age, tumor size, and levels of squamous cell carcinoma antigen (SCC), CA125, and CEA, as well as a significant difference in the CA19-9 level between MCT and squamous cell carcinoma arising from MCT, found that (1) age and tumor size are important factors in making a differential diagnosis and the optimal cutoff values for age and tumor size were, respectively, 45 years and 99 mm, (2)CEA was the best screening marker for squamous cell carcinoma arising from MCT, whereas age and tumor size were better markers than CA125 or CA19-9, and (3)SCC and CEA levels should be measured in patients age 45 years or older who have an MCT-like ovarian tumor larger than 99 mm in greatest dimension[1]. Robbins et al.(2009) examined the prognostic influence of reproductive factors on survival after ovarian cancer diagnosis,applied the Kaplan Meier method to estimate survival probability and Cox proportional hazard model to estimate risk ratio (HR) to study the prognostic impact of reproductive factors on survival rate after diagnosis of ovarian cancer, and found that high lifetime ovulatory cycles(LOC) and early age at menarche were associated with decreased survival after ovarian cancer[2]. Díaz-Padilla et al.(2012) summarized the clinical relevance of dynamic changes in CA-125 levels during the primary treatment of EOC and its potential influence both in the patient management and in the design of clinical trials in the adjuvant setting[3]. Anton et al.(2012) considered the contributions of the tumor markers CA125 and human epididymis protein 4 (HE4) as well as the risk ovarian

malignancy algorithm(ROMA) and risk malignancy index (RMI) values, evaluated their utility for establishing this system for patient referrals, and showed that there were no differences in accuracy between CA125, HE4, ROMA, and RMI for differentiating between types of ovarian masses[4]. Wang et al.(2014) conducted a meta-analysis to evaluate the diagnostic value of CA125, HE4 and ROMA in the diagnosis of ovarian cancer: systematically searched the PubMed and ScienceDirect databases and identified 32 studies that evaluated the role of CA125, HE4 and ROMA in diagnosing OCa, and observed that CA125 had a low specificity in premenopause and its diagnostic value was significantly lower than that postmenopausal, HE4 has higher specificity in premenopausal populations and can effectively reduce the misdiagnosis rate. Therefore, HE4 is suitable for diagnosing OC in premenopausal population, whereas CA125 and Roma are more suitable for the diagnosis of OC in postmenopausal population[5]. Muinao, Boruah & Pal(2019) discussed current trends in diagnostic approaches and updated potential several panels of cancer biomarkers for early detection of ovarian cancer, reported that CA125 in combinations with two or more biomarkers have outperformed single biomarker assays for early detection of the disease, found that CA-125 with CA 19 – 9, EGFR, G-CSF, Eotaxin, IL-2R, cVCAM, MIF improved the sensitivity with 98.2% and specificity of 98.7% in early stage detection of ovarian cancer, and demonstrated a panel of biomarkers signature as the potential tool for prototype development in future and other advanced approaches for early diagnosis of ovarian cancer to avoid false-diagnosis and excessive cost[6]. Lu et al.(2020) proposed machine learning to predict ovarian cancer, combined ovarian malignant tumor riskalgorithm (ROMA) with logistic regression model to accurately classify benign ovarian tumors (BOT) and Ovarian cancer (OC), and demonstrated that the machine learning approach had good potential in predictive modeling for the complex diseases, among others[7].

Logistic regression classifier is often used in disease diagnosis and finance forecasting. For high dimension classification problem, group penalized logistic regression classifier can improve the classification prediction performance by introducing group penalties to logistic regression. For example, Wei & Zhu(2012) studied group coordinate descent algorithms for nonconvex penalized regression including group SACD/MCP penalized logistic regression, showed that the estimated parameters converged to a global minimum when the sample size was larger than the dimension of the covariates, and converged to a local minimum otherwise, and found the group selection results of the MCP based and SCAD based GCD algorithms are better than the results selected by the group Lasso in terms of residual sum of squares and correct selection percentage[8]. Simon, Friedman & Hasti (2013) proposed a block coordinate descent algorithm to fit group penalized multiresponse and multinomial LASSO models, and verified the effectiveness of the algorithm[9]. Vincent & Hansen(2014) studied the classification performance from the multinomial sparse group LASSO based on a coordinate gradient descent algorithm, and showed that the multinomial group LASSO was significantly superior to the multinomial LASSO in terms of classification error rate and feature selection[10]. Sashimi et al.(2015) used semantic and phonological verbal fluency functional magnetic resonance imaging (fMRI) data to make a probabilistic diagnosis of depression, compared the performances of group lasso (gLASSO) and sparse group LASSO (sgLASSO) with standard LASSO(sLASSO),

SVM and random forest, demonstrated that gLASSO and sgLASSO outperformed sLASSO in classification robustness, identification of relevant brain regions and probability prediction were better than commonly used SVM and random forest[11]. Chen & Xiang (2017) constructed a credit scoring model based on group LASSO logistic regression, established three group LASSO models by AIC, BIC and cross-validation, and showed that the group LASSO method is superior to the backward elimination method in terms of interpretability and predictive accuracy[12]. Liu et al(2017) proposed a multi task sparsity group LASSO(MT-SGL) framework that deal with the loss function of generalized linear model to predict whether subjects are normal, mild cognitive impairment, or Alzheimer's disease[13]. Ghosal et al.(2020) proposed a new variable selection method in function linear concurrent regression, extended penalized variable selection methods (such as group LASSO, group SCAD and group MCP), and showed that the proposed method with group SCAD/MCP could pick out the relevant variables with high accuracy and had minuscule false positive and false negative rate even when data were observed sparsely, are contaminated with noise and the error process is highly non-stationary[14]. In this paper, we combine group LASSO/SCAD/MCP with logistic regression, and propose group LASSO/SCAD/MCP penalized logistic regression classifier to investigate the benign and malignant ovarian cancer. Firstly, we select the 46 predictor variables like blood routine, general chemical detection, tumor markers and basic information etc., and divide the selected 349 ovarian cancer patients into the two sets: the training set for learning and the testing set for predicting. We develop the group coordinate descent algorithm and the training samples to obtain group LASSO/SCAD/MCP estimator, and apply the testing samples to establish two-class confusion matrix, prediction accuracy, sensitivity and specificity, draw the ROC curve and apply the area under ROC curve (AUC) to assess the prediction performance. Finally, we compare group LASSO/SCAD/MCP penalized logistic regressions with SVM and ANN, and found that the prediction accuracy and AUC for group MCP/SCAD/LASSO penalized logistic regression/SVM/ANN is 93.33%/85.71%/82.26%/74.29%/72.38% and 0.892/0.852/0.823/0.639/0.789, respectively. So group MCP penalized logistic regressions performs the best.

The rest is arranged as follows: Section 2 specifies data source, 49 features and their group processing. Section 3 constructs three group penalized methods. Section 4 reports model estimators and the prediction performances for the five methods. Section 5 is conclusion.

2 Data and Features

2.1 Data source

Here we choose the 349 ovarian cancer patients composed of 178 benign ovarian tumor and 171 ovarian cancer from the Third Affiliated Hospital of Suzhou University from July 2011 to July 2018 selected from the kaggle website (https://www.kaggle.com/saurabhshahane/predict-ovarian-cancer). The data set is divide into two parts: a training set composed of 70% ovarian cancer patients and a testing set composed of 30% ovarian cancer patients, see Table 1.

Table 1. The specific ovarian cancer patients

	Benign ovarian tumor ($Y = 0$)	**Ovarian cancer** ($Y = 1$)	**Total**
Training set	98	146	n_1=244
Test set	80	25	n_2=105
Total	178	171	n=349

The chosen data set include 49 predictor variables: 19 blood routine tests, 22 general chemical tests, 6 tumor markers, age and menopause listed in Table 2. All patients experienced case diagnosis after operation, and none of them received preoperative radiotherapy and chemotherapy, and the histological type of diagnosis was classified according to the criteria of the World Health Organization.

Table 2. The 49 predictor variables and the response variable

Notation	**Variables**	**Definition**	**Value range**
X_1	MPV	Mean platelet volume	$7.4 \sim 12.5$(fL)
X_2	PLT	Platelet count	$125 \sim 350$(10^9/L)
X_3	PDW	Platelet distribution width	$15.5 \sim 18.1$(%)
X_4	PCT	Thrombocytocrit	$0.114 \sim 0.282$(L/L)
X_5	BASO#	Basophil cell count	$0 \sim 0.06$(10^9/L)
X_6	BASO%	Basophil cell ratio	$0 \sim 1$(%)
X_7	EO#	Eosinophil count	$0.02 \sim 0.52$(10^9/L)
X_8	EO%	Eosinophil ratio	$0.02 \sim 0.52$(%)
X_9	NEU	Neutrophil ratio	$40 \sim 75$(%)
X_{10}	LYM#	Lymphocyte count	$1.1 \sim 3.2$(10^9/L)
X_{11}	LYM%	Lymphocyte ratio	$20 \sim 50$(%)
X_{12}	MONO#	Mononuclear cell count	$0.1 \sim 0.6$(10^9/L)
X_{13}	MONO%	Monocyte ratio	$3 \sim 10$(%)
X_{14}	MCV	Mean corpuscular volume	$82 \sim 100$(fL)
X_{15}	MCH	Mean corpuscular hemoglubin	$27 \sim 34$(Pg)
X_{16}	RDW	Red blood cell distribution width	$10.6 \sim 15.5$(%)
X_{17}	HGB	Hemoglobin	$110 \sim 150$(g/L)
X_{18}	RBC	Red blood cell count	$3.5 \sim 5.5$(10^12/L)
X_{19}	HCT	Hematocrit	$0.35 \sim 0.45$(L/L)
X_{20}	Mg	Magnesium	$0.73 \sim 1.3$(mmol/L)
X_{21}	PHOS	Phosphorus	$0.7 \sim 1.62$(mmol/L)
X_{22}	Ca	Calcium	$1.12 \sim 1.32$(mmol/L)
X_{23}	Na	Natrium	$137 \sim 147$(mmol/L)
X_{24}	K	Kalium	$3.5 \sim 5.3$(mmol/L)

Table 2. The 49 predictor variables and the response variable

Notation	Variables	Definition	Value range
X_{25}	CL	Chlorine	$99 \sim 110$(mmol/L)
X_{26}	ALB	Albumin	$35 \sim 55$(g/L)
X_{27}	TP	Total protein	$60 \sim 82$(g/L)
X_{28}	GLO	Globulin	$20 \sim 40$(g/L)
X_{29}	TBIL	Total bilirubin	$4 \sim 19$(μmol/L)
X_{30}	DBIL	Direct bilirubin	$1.5 \sim 7$(μmol/L)
X_{31}	IBIL	Indirect bilirubin	$2 \sim 15$(μmol/L)
X_{32}	GGT	Gama glutamyltransferasey	$3 \sim 73$(U/L)
X_{33}	ALP	Alkaline phosphatase	$25 \sim 130$(U/L)
X_{34}	AST	Aspartate aminotransferase	$6 \sim 40$(U/L)
X_{35}	ALT	Alanine aminotransferase	$1 \sim 45$(U/L)
X_{36}	CA125	Carbohydrate antigen 125	$0 \sim 35$(U/mL)
X_{37}	CA19-9	Carbohydrate antigen 19-9	$0 \sim 37$(U/mL)
X_{38}	CA72-4	Carbohydrate antigen 72-4	$0 \sim 7$(U/mL)
X_{39}	AFP	Alpha-fetoprotein	$0 \sim 7$(ng/mL)
X_{40}	HE4	Human epididymis protein 4	$0 \sim 140$(pmol/L)
X_{41}	CEA	Carcinoembryonic antigen	$0 \sim 5$(ng/mL)
X_{42}	CREA	Creatinine	$44 \sim 144$(μmol/L)
X_{43}	UA	Urie acid	$90 \sim 450$(μmol/L)
X_{44}	BUN	Blood urea nitrogen	$1.7 \sim 8.3$(mmol/L)
X_{45}	AG	Anion gap	$8 \sim 30$(mmo/L)
X_{46}	CO2CP	Carban dioxide-combining power	$18 \sim 30$(mmol/L)
X_{47}	GLU	Glucose	$3.9 \sim 6.1$(mmol/L)
X_{48}	Age	Age	$15 \sim 83$
X_{49}	Menopause	Menopause	Premenopausal=0, Postmenopausal=1
Y	TYPE	Tumor type	BOT=0,OC=1

2.2 11 variable groups

In the raw data, total protein is the sum of albumin and globulin, total bilirubin is the sum of direct bilirubin and indirect bilirubin. Therefore, total protein and total bilirubin are deleted from 49 predictor variables. We observe that CA72-4 had 69% missing values. For convenience, we remove this variable. Therefore, total protein (X_{27}), total bilirubin (X_{29}) and carbohydrate antigen 72-4 (X_{38}) are removed, and the remaining 46 predictor variables with small missing values are filled with the mean value (the means of the missing variables is calculate separately

by tumor type and fill by tumor type). According to check item, we divide the remaining 46 predictor variables into 11 different variable groups, where blood routine test usually included relevant factors of platelet, white blood cell and red blood cell, liver function and renal function test are the basic tests of physiological functions of liver and kidney, and chemical, acid-base balance and blood sugar are used to measure whether the human body is in balanced state, and tumor markers are used to detect ovarian cancer. In the following Table 3 specifies the 11 variable groups.

Table 3. 11 variable groups

Group	Check item	Variables	Description
Group 1	Platelet	X_1, X_2, X_3, X_4	The main function of platelets is to accelerate coagulation, promote hemostasis and repair damaged blood vessels.
Group 2	White blood cell	$X_5, X_6, X_7, X_8, X_9, X_{10}, X_{11}, X_{12}, X_{13}$	White blood cells can phagocytose foreign materials to produce antibodies, and heal body damage, resist pathogen invasion and disease immunity.
Group 3	Red blood cell	$X_{14}, X_{15}, X_{16}, X_{17}, X_{18}, X_{19}$	The main work of red blood cells is to transport oxygen and carbon dioxide which can enhance phagocytosis and immune adhesion.
Group 4	Chemical element	$X_{20}, X_{21}, X_{22}, X_{23}, X_{24}, X_{25}$	Ions are used to measure human body electrolytes. The imbalance of the number of cations and anions will cause electrolyte disorders,which will lead different body damages.
Group 5	Liver function	$X_{26}, X_{28}, X_{30}, X_{31}, X_{32}, X_{33}, X_{34}, X_{35}$	Liver function examination generally includes protein metabolism function, bilirubin and bile acid metabolism function and serum enzyme indexes.

Table 3. 11 variable groups

Group	Check item	Variables	Description
Group 6	Tumor marker	$X_{36}, X_{37}, X_{39}, X_{40}, X_{41}$	Tumor markers can be used for early detection, screening and differential diagnosis of tumors and can also be used for patient efficacy detection, recurrence and prognosis judgment.
Group 7	Renal function	X_{42}, X_{43}, X_{44}	The main function of kidney is to secrete and excrete urine and toxins, regulate body fluids volume and water, and maintain the balance of body's internal environment.
Group 8	Acid-base balance	X_{45}, X_{46}	The pH value of normal people's blood is always maintained at a certain level. Once the acid-base balance is disturbed, acidosis or alkalosis will occur.
Group 9	Blood sugar	X_{47}	The glucose in the blood is called blood sugar. The production and utilization of blood sugar are in a state of dynamic balance to maintain the needs of various organs and tissues in the body.
Group 10	Age	X_{48}	Ovarian cancer has a certain relationship with age. The most common age group for ovarian cancer is middle-aged and elderly women, but many young women may also suffer from ovarian cancer.
Group 11	Menopause	X_{49}	Early and late amenorrhea have a certain impact on women's physical health. The later the amenorrhea, the greater the risk of ovarian cancer.

3 The three group penalized methods

3.1 Group LASSO penalized logistic regression

Tibshirani(1996) firstly introduced L_1 function to linear model, and proposed LASSO (Least Absolute Shrink and Selection Operator) penalized linear model for variable selection[15]. Then,

LASSO penalized logistic regression, group LASSO penalized linear regression and group LASSO penalized logistic regression are proposed. For example, Yuan & Lin (2006) [16]proposed group LASSO penalized linear regression and constructed its log likelihood function

$$Q(\beta;\lambda)=\frac{1}{2}\left\|Y-\sum_{j=1}^{J}X_{(j)}\beta_{(j)}\right\|_2^2+\lambda\sum_{j=1}^{J}\sqrt{d_j}\|\ \beta_{(j)}\|_2\,, \tag{1}$$

where $\beta=(\beta_{(1)},\ldots,\beta_{(J)})$ with the jth group coefficient vector

$$\beta_{(j)}=(\beta_{d_{j-1}+1},\ldots,\beta_{d_{j-1}+d_j})^\top \tag{2}$$

is the whole coefficient vector, and $d_j=\dim(\beta_{(j)})$ is the length of the j-th group. The j-th group LASSO estimator is given by

$$\widehat{\beta}_{(j)}=\left(1-\frac{\lambda\sqrt{d_j}}{\|\ S_j\ \|}\right)_+S_j, j=1,\ldots,J, \tag{3}$$

where $S_j=X_{(j)}^\top(Y-\beta_{-(j)})$ and $\beta_{-(j)}=(\beta_{(1)}^\top,\ldots,\beta_{(j-1)}^\top,0,\beta_{(j+1)}^\top,\ldots,\beta_{(J)}^\top)$. Meier, van de Geer & Bühlmann (2008) combined group LASSO penalty with logistic regression and proposed group LASSO penalized logistic regression (GLASSO) to investigate ovarian cancer[17]. In this paper we introduce group logistic regression

$$\log\frac{P_\beta\left(X_i\right)}{1-P_\beta\left(X_i\right)}=\beta_0+\sum_{j=1}^{11}X_{i(j)}^\top\beta_{(j)}=\eta_i, i=1,\ldots,244, j=1,\cdots,11, \tag{4}$$

to study the relation between Y and $X=(X_{(1)},\ldots,X_{(11)})^\top$, where

$$P_\beta\left(X_i\right)=P\left(Y=1\mid X_i;\beta\right)=\frac{\exp\left(\beta_0+\sum\limits_{j=1}^{11}X_{i(j)}^\top\beta_{(j)}\right)}{1+\exp\left(\beta_0+\sum\limits_{j=1}^{11}X_{i(j)}^\top\beta_{(j)}\right)}$$

is the conditional probability of ovarian benign tumors, β_0 is the intercept, $\beta_{(j)}$ is the j-th group parameter vector and $\beta=(\beta_{(1)},\ldots,\beta_{(11)})$ is the whole unknown parameter vector. Then, the negative group log likelihood function for group logistic regression is

$$\begin{aligned}L(\beta)=-l(\beta)&=-\frac{1}{244}\sum_{i=1}^{244}\left\{Y_i\log P_\beta\left(X_i\right)+\left(1-Y_i\right)\log\left(1-P_\beta\left(X_i\right)\right)\right\}\\&=-\frac{1}{244}\sum_{i=1}^{244}\left\{Y_i\left(\beta_0+\sum_{j=1}^{11}X_{i(j)}^\top\beta_{(j)}\right)-\log\left[1+\exp\left(\beta_0+\sum_{j=1}^{11}X_{i(j)}^\top\beta_{(j)}\right)\right]\right\}.\end{aligned} \tag{5}$$

Group LASSO penalized logistic log likelihood is

$$Q(\beta;\lambda) = L(\beta) + \lambda \sum_{j=1}^{11} \sqrt{d_j} \left\| \beta_{(j)} \right\|, \tag{6}$$

where the tuning parameter $\lambda \geq 0$ controls the penalty size. Suppose that for a univariate Z, the univariate soft-thresholding operator is

$$S(Z,\lambda) = \begin{cases} Z - \lambda, & \text{if } Z > \lambda, \\ 0, & \text{if } |Z| \leq \lambda, \\ Z + \lambda, & \text{if } Z < -\lambda, \end{cases} \tag{7}$$

and for a vector-valued argument Z, the multivariate soft-thresholding operator is

$$S(Z,\lambda) = S(\|Z\|,\lambda)\frac{Z}{\|Z\|}, \tag{8}$$

where $Z/\|Z\|$ is the unit vector in the direction of Z. In other words, $S(Z,\lambda)$ acts on the vector Z by shortening it towards to 0, and if the length of Z is less than λ, the vector is shortened all the way to 0. Let $\eta = \beta_0 + \sum_{j=1}^{11} X_{(j)}^{\top}\beta_{(j)}, \eta_i = \beta_0 + \sum_{j=1}^{11} X_{i(j)}^{\top}\beta_{(j)},\ v = \max_i \sup_\eta \{\nabla^2 L_i(\eta)\}$ with $L_i(\eta) = Y_i\eta_i - \log\left(1 + e^{\eta_i}\right), i = 1, \ldots, 244$, so that $vI - \nabla^2 L(\eta)$ is positive semi-definite matrix at all points η. Breheny & Huang (2011,2015)[18,19] proposed group coordinate descent algorithm for group LASSO penalized logistic regression(GLASSO-PLR) as follows:

Algorithm 1 Group coordinate descent algorithm for GLASSO-PLR

1. Let $\eta = \beta_0 + \sum_{j=1}^{11} X_{(j)}^{\top}\beta_{(j)}, \eta_i = \beta_0 + \sum_{j=1}^{11} X_{i(j)}^{\top}\beta_{(j)}, i = 1, \cdots, 244,$
   $P = e^{\eta}/\left(1 + e^{\eta}\right), L_i(\eta) = Y_i\eta_i - \log\left(1 + e^{\eta_i}\right), v = \max_i \sup_\eta \{\nabla^2 L_i(\eta)\}$;
2. At the j-th step of $m+1$ iterations, $j = 1, 2, \cdots, 11$, carry out the following (A)-(C):
   (A) compute $\tilde{r} \leftarrow (Y - P)/v$ and $Z_{(j)} = X_{(j)}^{\top}\tilde{r} + \beta_{(j)}$;
   (B) update $\hat{\beta}_{(j)}^{GLASSO}(m+1) \leftarrow S\left(vZ_{(j)}, \lambda_j\right)/v$;
   (C) update residual $\tilde{r}^{\top} \leftarrow \tilde{r} - X_{(j)}^{\top}\left(\hat{\beta}_{(j)}^{GLASSO}(m+1) - \hat{\beta}_{(j)}^{GLASSO}(m)\right)$;
3. Update $m \leftarrow m + 1$;
4. Repeat 2 and 3 until convergence.

Introduce $X_{-(j)} = \left(X_{(1)}^{\top}, \cdots, X_{(j-1)}^{\top}, 0^{\top}, X_{(j+1)}^{\top}, \cdots, X_{(11)}^{\top}\right)$, $P_i = e^{\eta_i}/\left(1 + e^{\eta_i}\right)$,

$$\beta_{-(j)} = \left(\beta_{(1)}^{\top}, \cdots, \beta_{(j-1)}^{\top}, 0^{\top}, \beta_{(j+1)}^{\top}, \cdots, \beta_{(11)}^{\top}\right), W = diag\{P_i(1 - P_i)\}.$$

Compute $\tilde{Y} = X^{\top}\hat{\beta}^{GLASSO}(m) + W^{-1}(Y - P)$ and $Z_{(j)} = X_{(j)}^{\top}\left(\tilde{Y} - X_{-(j)}\beta_{-(j)}\right)$, We call the grprep package and obtain the group LASSO estimator

$$\hat{\beta}_{(j)}^{GLASSO} = \frac{1}{v}S\left(vZ_{(j)}, \lambda\sqrt{d_j}\right) = \frac{1}{v}S\left(v\|Z_{(j)}\|, \lambda\sqrt{d_j}\right)\frac{Z_{(j)}}{\left\|Z_{(j)}\right\|}, j = 1, \ldots, 11. \tag{9}$$

3.2 Group SCAD penalized logistic regression

Fan & Li(2001)[20] proposed the following SCAD penalty

$$P_{SCAD}(\beta;\lambda,\gamma)=\begin{cases}\lambda\beta, & \text{if } \beta\le\lambda,\\ \frac{2\lambda\gamma\beta-\left(\beta^2+\lambda^2\right)}{2(\gamma-1)}, & \text{if } \lambda<\beta\le\gamma\lambda,\\ \frac{\lambda^2\left(\gamma^2-1\right)}{2(\gamma-1)}, & \text{if } \beta>\gamma\lambda.\end{cases} \tag{10}$$

where $\lambda>0$ and $\gamma>2$. Its first derivative with respect to the parameter vector β is

$$P^{'}_{SCAD}(\beta;\lambda,\gamma)=\lambda\left\{I(\beta\le\lambda)+\frac{(\gamma\lambda-\beta)_+}{(\gamma-1)\lambda}I(\beta>\lambda)\right\}=\begin{cases}\lambda, & \text{if } \beta\le\lambda,\\ \frac{(\gamma\lambda-\beta)}{\gamma-1}, & \text{if } \lambda<\beta\le\lambda\gamma,\\ 0, & \text{if } \beta>\lambda\gamma.\end{cases} \tag{11}$$

Group penalized log-likelihood for group SCAD penalized logistic regression(GSCAD-PLR) is

$$Q(\beta;\lambda,\gamma)=L(\beta)+\sum_{j=1}^{11}P_{SCAD}\left(\left\|\beta_{(j)}\right\|;\lambda\sqrt{d_j},\gamma\right). \tag{12}$$

Similar to Algorithm 1, we apply the group coordinate descent algorithm for GSCAD-PLR and obtain the j-th group SCAD estimator

$$\hat{\beta}^{GSCAD}_{(j)}=\begin{cases}\frac{1}{v}S\left(vZ_{(j)},\lambda\sqrt{d_j}\right), & \text{if } \|Z_{(j)}\|\le 2\lambda\sqrt{d_j},\\ \frac{\gamma-1}{\gamma-2}\cdot\frac{1}{v}S\left(vZ_{(j)},\frac{\lambda\sqrt{d_j}\gamma}{\gamma-1}\right), & \text{if } 2\lambda\sqrt{d_j}<\left\|Z_{(j)}\right\|\le\lambda\sqrt{d_j}\gamma,\gamma>2,\\ Z_{(j)}, & \text{if } \parallel Z_{(j)}\parallel>\lambda\sqrt{d_j}\gamma.\end{cases} \tag{13}$$

3.3 Group MCP penalized logistic regression

Zhang (2010)[21] proposed the nonconvex penalized function MCP(minimax convex penalty)

$$P_{MCP}(\beta;\lambda,\gamma)=\begin{cases}\lambda\beta-\frac{\beta^2}{2\gamma}, & \text{if } \beta\le\gamma\lambda,\\ \frac{\gamma\lambda^2}{2}, & \text{if}\beta>\gamma\lambda.\end{cases} \tag{14}$$

where $\lambda>0$ and $\gamma>1$. Its first derivative with respect to the parameter vector β is

$$P^{'}_{MCP}(\beta;\lambda,\gamma)=\lambda\left(1-\frac{\beta}{\lambda\gamma}\right)_+\operatorname{sgn}(\beta)=\begin{cases}\lambda-\frac{\beta}{\gamma}, & \text{if } \beta\le\lambda\gamma,\\ 0, & \text{if } \beta>\lambda\gamma.\end{cases} \tag{15}$$

where $\operatorname{sgn}(\beta)=-1,0,1$ corresponds to $\beta<0,\beta=0,\beta>0$, respectively. The group penalized log-likelihood for group MCP penalized logistic regression(GMCP-PLR) is

$$Q(\beta;\lambda,\gamma)=L(\beta)+\sum_{j=1}^{11}P_{MCP}\left(\left\|\beta_{(j)}\right\|;\lambda\sqrt{d_j},\gamma\right). \tag{16}$$

Similar to Algorithm 1, we apply the group coordinate descent algorithm for GMCP-PLR and obtain the j-th group MCP estimator

$$\hat{\beta}_{(j)}^{GMCP} = \begin{cases} \frac{\gamma}{\gamma-1} \cdot \frac{1}{v} S\left(vZ_{(j)}, \sqrt{d_j}\lambda\right), & \text{if } \| Z_{(j)}\| \leq \sqrt{d_j}\lambda\gamma, \gamma > 1, \\ Z_{(j)}, & \text{if } \left\|Z_{(j)}\right\| > \sqrt{d_j}\lambda\gamma. \end{cases} \tag{17}$$

3.4 Two-class prediction accuracy evaluation

For two-class problem, a two-class confusion matrix(accuracy, sensitivity, and specificity), a ROC curve and the area under the ROC curve(AUC) are often used as the prediction performance evaluation indexes. Table 4 lists the two-class confusion matrix.

Table 4. Two class confusion matrix

	True Class 1	**True Class 2**
Prediction Class 1	TP(True Positives)	FP(False Positives)
Prediction Class 2	FN(False Negatives)	TN(True Negatives)

From table 4, one can compute

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FN+FP}, \quad \text{Precision} = \frac{TP}{TP+FP},$$
$$\text{Sensitivity} = \frac{TP}{TP+FN}, \quad \text{Specificity} = \frac{TN}{TN+FP}.$$

The ROC curve can be drawn by changing $(1-$ specificity) and sensitivity at different thresholds. $(1-$ specificity) is x-axis that represents false positive rate (FPR). The smaller the FPR, the lower the false positive rate, and the less the actual negative class in the predicted negative class. Sensitivity is y-axis that represents true positive rate (true postive rate, TPR). The larger the TPR, the higher the hit, the more the actual positive class in the predicted positive class. The larger the AUC, the better the prediction performance. AUC ranges from 0 to 1: AUC$\in (0, 0.5)$ reflects the worse prediction performance than random guess; AUC $= 0.5$ reflects the bad prediction performance like random guess; AUC$\in (0.5, 0.7)$ reflects the low prediction performance, whereas for stock prediction, AUC$\in (0.5, 0.7)$ reflects the good prediction performance; AUC$\in (0.7, 0.9)$ reflects the relative high prediction accuracy; AUC $\in (0.9, 1)$ reflects the very high prediction accuracy and AUC=1 reflects the perfect prediction performance.

3.5 Path selection

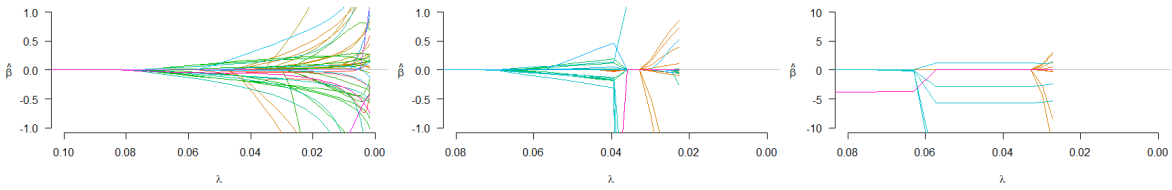
The tuning parameter λ controls the size of the penalized strength. The larger λ, the stronger the penalized degree, the more coefficients are compressed to 0, and the less non-zero parameters are chosen. Therefore, the choice of the tuning parameter λ is crucial. Commonly used methods are Akaike information criterion (AIC), Bayesian information criterion (BIC) and cross validation (CV). However, we are interested in obtaining $\hat{\beta}$ not just for a single value of λ, but for a range of values extending from a maximum value $\lambda_{\max}$ for which all penalized

coefficients are 0 down to $\lambda = 0$ or to a minimum value $\lambda_{\min}$ at which the model becomes excessively large or ceases to be identifiable. Here, we consider λ on a grid $\{\lambda_0, \ldots, \lambda_{K+1}\}$, Then, we apply ten cross validation to select the optimal λ, and obtain $\hat{\beta}$ through the optimal λ. For default γ, group SCAD is 4 and group MCP is 3, the algorithm starts at $\lambda_{\max}$ and proceeding toward $\lambda_{\min}$. When the objective function is a strictly convex function, the estimated coefficients continuously vary within $\lambda \in [\lambda_{\min}, \lambda_{\max}]$ and produce a regularized solution path. Algorithm 1 is an iterative algorithm, the maximum $\lambda_{\max}$ at $\beta = 0$ determined as the iterative initial value, $\lambda_{\max} = \max_j\{v\|Z_{(j)}\|\}$ for logistic regression, and going from the maximum toward to the minimum, $\hat{\beta}$ obtained by the previous λ as the initial value for the next one to ensure that the initial value does not break away from solution.

4 Model estimators and prediction performances

4.1 Model estimators

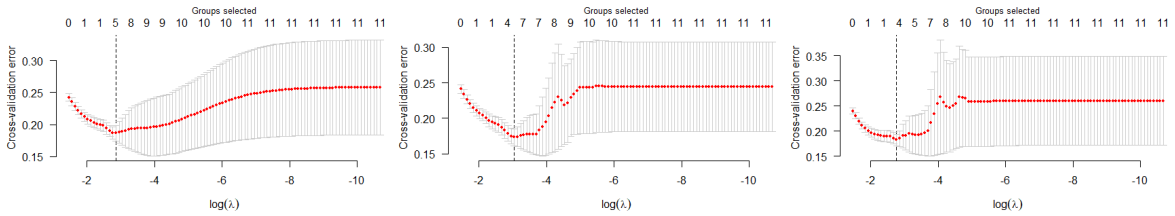
Group LASSO/SCAD/MCP estimators for group penalized logistic regressions are obtained by the grpreg package. Firstly, the training set is orthogonalised within the group, the optimal λ is selected by ten fold cross validation, and apply the optimal λ and default γ and the formula (9),(13)and (17) to compute group LASSO/SCAD/MCP estimators. Then, the test set is used to compute a confusion matrix, accuracy, sensitivity, specificity and draw ROC curves so that one can compare the prediction accuracy. Fig.1 shows the coefficient path diagrams selected by group LASSO/SCAD/MCP penalty, where the abscissa is the λ value and the ordinate is the coefficient estimators $\widehat{\beta}$.



(a) Group LASSO (b) Group SCAD (c) Group MCP

Fig. 1. The coefficient path diagrams for group LASSO/SCAD/MCP.

The optimal λ selected by ten fold cross validation are listed in Fig.2. Ordinates represent cross validation errors, abscissa represents $\log(\lambda)$ and the numbers above indicate the number of variables entered into the model at the corresponding λ value. Table 5 lists the optimal λ selected by ten fold cross validation based on ovarian tumors data sets for group LASSO/SCAD/MCP penalty.



(a) Group LASSO (b) Group SCAD (c) Group MCP

Fig. 2. The cross validation errors diagrams for group LASSO/SCAD/MCP.

Table 5. Optimal lambda values for group LASSO/SCAD/MCP

	Group LASSO	**Group SCAD**	**Group MCP**
λ value	0.0462	0.0430	0.0520

From Table 5, we found that the optimal λ of group MCP penalized logistic regression is larger than group LASSO/SCAD penalized logistic regressions. Therefore, the penalization intensity of group MCP is greater, the more coefficients compressed to 0, and the fewer variable groups selected. After determining the optimal λ selected by ten fold cross validation, we apply the group coordinate descent algorithm to obtain group estimators for group LASSO/SCAD/MCP penalized logistic regressions. The group coefficient estimators selected by group LASSO/SCAD/MCP penalty are listed in Table 6.

Table 6. Group LASSO/SCAD/MCP estimators

Variable	**Group LASSO**	**Group SCAD**	**Group MCP**	**Variable**	**Group LASSO**	**Group SCAD**	**Group MCP**
Intercept	-1.7058	-2.6146	-4.8250	$\beta_{(4-24)}$	0	0	0
$\beta_{(1-1)}$	-0.0781	0	0	$\beta_{(4-25)}$	0	0	0
$\beta_{(1-2)}$	-0.0238	0	0	$\beta_{(5-26)}$	-0.0830	-0.1234	0
$\beta_{(1-3)}$	-0.0213	0	0	$\beta_{(5-28)}$	-0.1639	-0.1591	0
$\beta_{(1-4)}$	0.0104	0	0	$\beta_{(5-30)}$	0.0370	0.0395	0
$\beta_{(2-5)}$	0	0	0	$\beta_{(5-31)}$	0.1349	0.1628	0
$\beta_{(2-6)}$	0	0	0	$\beta_{(5-32)}$	0.0061	0.0180	0
$\beta_{(2-7)}$	0	0	0	$\beta_{(5-33)}$	0.1263	0.1025	0
$\beta_{(2-8)}$	0	0	0	$\beta_{(5-34)}$	-0.1672	-0.1532	0
$\beta_{(2-9)}$	0	0	0	$\beta_{(5-35)}$	-0.1691	-0.1593	0
$\beta_{(2-10)}$	0	0	0	$\beta_{(6-36)}$	-0.2574	-0.2659	-17.3783
$\beta_{(2-11)}$	0	0	0	$\beta_{(6-37)}$	0.0988	0.1180	1.1620
$\beta_{(2-12)}$	0	0	0	$\beta_{(6-39)}$	-0.0903	-0.1089	-5.6947
$\beta_{(2-13)}$	0	0	0	$\beta_{(6-40)}$	-0.1336	-0.1411	-20.3392
$\beta_{(3-14)}$	0	0	0	$\beta_{(6-41)}$	-0.0942	-0.1122	-2.8399
$\beta_{(3-15)}$	0	0	0	$\beta_{(7-42)}$	0.1868	0.3735	0
$\beta_{(3-16)}$	0	0	0	$\beta_{(7-43)}$	0.0229	0.0388	0

Table 6. Group LASSO/SCAD/MCP estimators

Variable	Group LASSO	Group SCAD	Group MCP	Variable	Group LASSO	Group SCAD	Group MCP
$\beta_{(3-17)}$	0	0	0	$\beta_{(7-44)}$	-0.0045	0.0003	0
$\beta_{(3-18)}$	0	0	0	$\beta_{(8-45)}$	0	0	0
$\beta_{(3-19)}$	0	0	0	$\beta_{(8-46)}$	0	0	0
$\beta_{(4-20)}$	0	0	0	$\beta_{(9-47)}$	0	0	0
$\beta_{(4-21)}$	0	0	0	$\beta_{(10-48)}$	-2.1425	-3.3630	0
$\beta_{(4-22)}$	0	0	0	$\beta_{(11-49)}$	-0.0528	0	0
$\beta_{(4-23)}$	0	0	0				

From Table 6, we can see that group LASSO penalized logistic regression retains the 6 groups composed of 22 non-zero explanatory variables in 6 groups and compresses the other 5 groups composed of 24 explanatory variables to 0. Group SCAD penalized logistic regression retains the 4 groups composed of 17 non-zero explanatory variables. Group MCP penalized logistic regression selected tumor marker group among 11 groups of explanatory variables, including CA125, CA19-9, AFP, HE4 and CEA, and the coefficients of the remaining explanatory variables are compressed to 0, indicating that the 10 variable groups have no significant influence on the discrimination of benign and malignant ovarian tumors. Therefore, group MCP penalized logistic regression can effectively predict the benign and malignant of ovarian tumors. Its probability estimators are as follows:

$$\hat{P}(Y_i = 1|X_i) = \frac{e^{-4.8250-17.3783X_{36}+1.1620X_{37}-5.6947X_{39}-20.3392X_{40}-2.8399X_{41}}}{1+e^{-4.8250-17.3783X_{36}+1.1620X_{37}-5.6947X_{39}-20.3392X_{40}-2.8399X_{41}}}.$$

$$\hat{P}(Y_i = 0|X_i) = \frac{1}{1+e^{-4.8250-17.3783X_{36}+1.1620X_{37}-5.6947X_{39}-20.3392X_{40}-2.8399X_{41}}}.$$

We conclude that the tumor marker group is the critical indicator for the diagnosis of benign and malignant ovarian tumors.

4.2 Prediction performances

In the following we apply group LASSO/SCAD/MCP penalized logistic regressions to predict benign and malignant ovarian tumors. Apply the test set $\{(X_i, Y_i), i = n_1+1, \ldots, n_1+n_2\}$ with $n_1 = 244$ and $n_2 = 105$ to the probability estimators, and estimate the predicted values $\widehat{Y}_i$ according to the following rules:

$$\text{If} \quad \hat{P}_i > c, \quad \text{then} \quad \widehat{Y}_i = 1, \quad \text{else} \quad \widehat{Y}_i = 0, \tag{18}$$

where c is a given threshold. For balanced data, c is generally taken as 0.5. For unbalanced data, Youden index is widely used to select the optimal threshold (Raghavan, Ashour & Bailey, 2016)[22].

To evaluate the prediction performance, we compare group LASSO/SCAD/MCP penalized logistic regressions with SVM and ANN. The confusion matrixes and the prediction performances from the five methods are listed in Table 7 and Table 8, respectively.

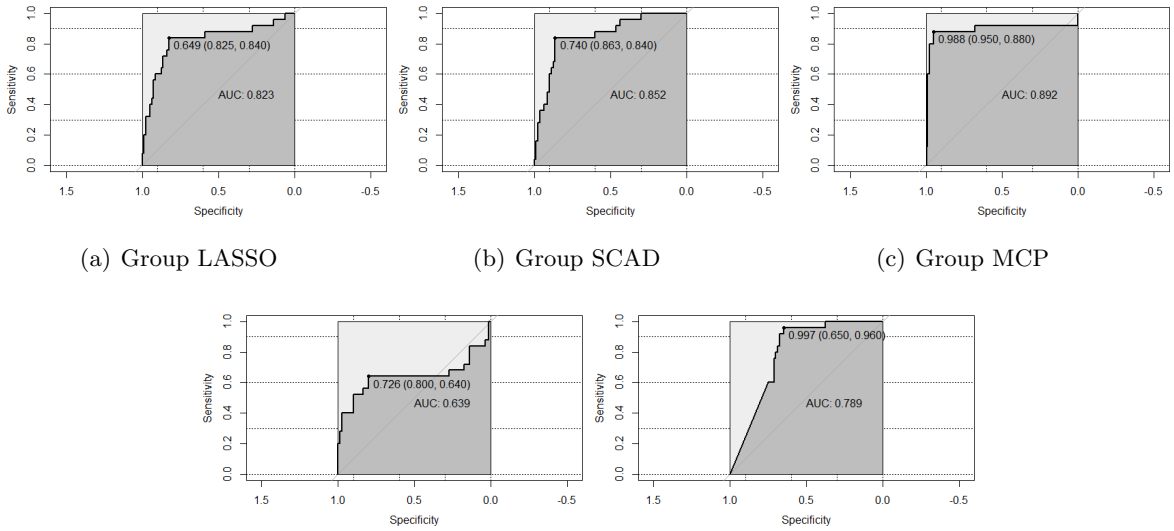
Table 7. Two-class confusion matrix comparisons

		1(OC)	**0(BOT)**
Group LASSO	1(OC)	21	14
	0(BOT)	4	66
Group SCAD	1(OC)	21	11
	0(BOT)	4	69
Group MCP	1(OC)	22	4
	0(BOT)	3	76
SVM	1(OC)	14	16
	0(BOT)	11	64
ANN	1(OC)	24	28
	0(BOT)	1	52

Table 8. The prediction performance comparisons

Model	**Accuracy**	**Precision**	**Specificity**	**Sensitivity**
Group LASSO	0.8286	0.6000	0.8250	0.8400
Group SCAD	0.8571	0.6563	0.8625	0.8400
Group MCP	0.9333	0.8462	0.9500	0.8800
SVM	0.7429	0.4667	0.8000	0.5600
ANN	0.7238	0.4615	0.6500	0.9600

According to Table 8, the prediction performance of group penalized logistic regression is better than that of machine learning method. The prediction accuracy, precision and specificity for group MCP penalized logistic regression are 93.33%, 84.62% and 95%, respectively. The prediction accuracy for group SCAD penalized logistic regression is 85.71%. The predictive accuracy for group LASSO penalized logistic regression is 82.86%. The prediction accuracy of both SVM and ANN are 74.29% and 72.38%, respectively. ANN predicts the worst with a prediction accuracy that differed by 20.95% from the group MCP penalized logistic regression, as well as the lowest precision and specificity with 46.15% and 65%, but the sensitivity is 96% that is the highest. The prediction performance for SVM is slightly better than that of ANN with a prediction accuracy of 74.29%, and its prediction sensitivity is 56% that is the lowest. The ROC curve and the area under the ROC curve(AUC) are often used to evaluate the predictive performance. The larger the AUC, the better the predictive performance. Here we apply pROC package to visualize their ROC curves, see Fig.3.



(a) Group LASSO (b) Group SCAD (c) Group MCP

(d) SVM (e) ANN

Fig. 3. The ROC curves for the five methods.

As shown in Fig.3, the sensitivity and specificity of several models are consistent with Table 8, the optimal thresholds for group LASSO/SCAD/MCP penalized logistic regressions are 0.649, 0.740 and 0.988, respectively, and the AUC values are 0.823, 0.852, 0.892, respectively. The optimal threshold for SVM is 0.726 and the AUC value is 0.639. The optimal threshold value for ANN is 0.997 and the AUC value is 0.789. We found that the AUC values for group penalized logistic regressions have exceeded 0.8 that reflects very high prediction accuracy. AUC values for both SVM and ANN are below 0.8 that reflects the relatively poor prediction performance. Thus, group MCP penalized logistic regression has the highest AUC value and the best prediction performance in benign and malignant ovarian tumors prediction problem.

5 Conclusion

In this paper we propose group LASSO/SCAD/MCP penalized logistic regressions to predict ovarian tumors. We select 46 explanatory variables and divide them into 11 variable groups, develop the three group penalized methods to select the significant variable groups and found that the group of tumor markers is the key variable group to predict the benign or malignant ovarian tumors, where the optimal tuning parameters are selected by ten folds cross-validation. Group LASSO penalized logistic regression retains 6 variable groups composed of 22 explanatory variables. Group SCAD penalized logistic regression retains 4 variable groups composed of 17 nonzero explanatory variables. Group MCP penalized logistic regression only selects 1 tumor marker group composed of 5 explanatory variables: CA125, CA19-9, AFP, HE4 and CEA. Finally, we compare the proposed group LASSO/SCAD/MCP penalized logistic regression with ANN and SVM, and found that the prediction accuracy and AUC for group MCP/SCAD/LASSO penal-

ized logistic regression/SVM/ANN is 93.33%/85.71%/82.26%/74.29%/72.38% and 0.892/0.852 /0.823/0.639/0.789, respectively. Obviously, the proposed group MCP penalized logistic regression performs the best. Therefore, we propose group MCP penalized logistic regression to predict ovarian cancer.

Acknowledgements This research was supported by the Fifth Batch of Excellent Talent Support Program of Chongqing Colleges and University (68021900601), the Natural Science Foundation of CQ CSTC (cstc.2018jcyjA2073), the Program for the Chongqing Statistics Postgraduate Supervisor Team (yds183002), Chongqing Social Science Plan Project (2019WT59), Science and Technology Research Program of Chongqing Education Commission(KJZD-M202100801), Open Project from Chongqing Key Laboratory of Social Economy and Applied Statistics (KFJJ2018066) and Mathematic and Statistics Team from Chongqing Technology and Business University (ZDPTTD201906).

Author contributions Xuemei Hu provided the basic idea, the important guidance and complete the final writing. Ying Xie collected data, wrote the program, provided figures and tables and finished the elementary writing. Yanlin Yang and Huifeng Jiang improved the program.

Funding This research was supported by the Fifth Batch of Excellent Talent Support Program of Chongqing Colleges and University (68021900601), the Natural Science Foundation of CQ CSTC (cstc.2018jcyjA2073), the Program for the Chongqing Statistics Postgraduate Supervisor Team (yds183002), Chongqing Social Science Plan Project (2019WT59), Science and Technology Research Program of Chongqing Education Commission(KJZD-M202100801), Open Project from Chongqing Key Laboratory of Social Economy and Applied Statistics (KFJJ2018066) and Mathematic and Statistics Team from Chongqing Technology and Business University (ZDPTTD201906).

Availability of data The data set used in this study are available from https://www.kaggle.com/saurabhshahane/predict-ovarian-cancer.

Declarations

Conflict of interest The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Ethic approval This is an observational study and do not require ethics approval.

Consent to publication All the authors consented to the publication of this article.

Reference

[1] Kikkawa F , Nawa A , Tamakoshi K , Ishikawa H, Kuzuya K, Suganuma N, Hattori S, Furui K, Kawai M, Arii Y. Diagnosis of squamous cell carcinoma arising from mature cystic teratoma of the ovary[J]. Cancer, 1998, 82(11):2249-2255.DOI:10.1002/(SICI)1097-0142(19980601)82:113.0.CO;2-T.

[2] Robbins C L, Whiteman M K, Hillis S D, Curtis K M, McDonald J A, Wingo P A, Kulkarni A, Marchbanks P A. Influence of reproductive factors on mortality after epithelial ovarian cancer diagnosis[J]. Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology, 2009,18(7):2035-2041.DOI:10.1158/1055-9965.EPI-09-0156.

[3] Díaz-Padilla I, Razak A R A, Minig L , Bernardini M Q, del Campo J M. Prognostic and predictive value of CA-125 in the primary treatment of epithelial ovarian cancer: potentials and pitfalls[J]. Clinical & Translational Oncology, 2012, 14(1):15-20.DOI:10.1007/s12094-012-0756-8.

[4] Anton C, Carvalho F M, Oliveira E I, Maciel G A R, Baracat E C, Carvalho J P. A comparison of CA125, HE4, risk ovarian malignancy algorithm (ROMA), and risk malignancy index (RMI) for the classification of ovarian masses[J]. Clinics (Sao Paulo), 2012, 67 (5)437-441.DOI:10.6061/clinics/2012(05)06.

[5] Wang J, Gao J, Yao H, Wu Z , Wang M, Qi J. Diagnostic accuracy of serum HE4, CA125 and ROMA in patients with ovarian cancer: a meta-analysis[J]. Tumor Biology, 2014, 35(6):6127-6138.DOI:10.1007/s13277-014-1811-6.

[6] Muinao T, Boruah H P D, Pal M. Multi-biomarker panel signature as the key to diagnosis of ovarian cancer[J]. Heliyon,2019,5(12),DOI:10.1016/j.heliyon.2019.e02826.DOI:10.1016/j.heliyon.2019.e02826.

[7] Lu M, Fan Z, Xu B, Chen L, Zheng X,Li J, Znati T, Mi Q, Jiang J. Using machine learning to predict ovarian cancer[J]. International Journal of Medical Informatics,2020,141,DOI:10.1016/j.ijmedinf.2020.104195.DOI:10.1016/j.ijmedinf.2020.104195.

[8] Wei F R, Zhu H X. Group coordinate descent algorithms for nonconvex penalized regression[J]. Computational Statistics and Data Analysis, 2012,56(2):316-326.DOI:10.1016/j.csda.2011.08.007.

[9] Simon N, Friedman J, Hastie T. A blockwise descent algorithm for group-penalized multiresponse multinomial regression[J]. Statistics,2013. Cite:arXiv:1311.6529.

[10] Vincent M, Hansen N R. Sparse group LASSO and high dimensional multinomial classification[J]. Computational Statistics and Data Analysis, 2014,71: 771-786.Cite:arXiv:1205.1245.

[11] Sashimi Y, Yoshimoto J, Toki S, Takamura M, Yoshimura S, Okamoto Y, Yamawaki S, Doya K. Toward probabilistic diagnosis and understanding of depression based on functional MRI data analysis with logistic group LASSO[J]. PLoS ONE, 2015,10(5).DOI:10.1371/journal.pone.0123524.

[12] Chen H, Xiang Y. The study of credit scoring model based on group LASSO[J]. Procedia Computer Science, 2017,122:677-684.DOI:10.1016/j.procs.2017.11.423.

[13] Liu X, Goncalves AR, Cao P, Zhao D, Banerjee A, Alzheimer's Disease Neuroimaging Initiative. Modeling Alzheimer's disease cognitive scores using multi-task sparse group LASSO[J]. Comput Med Imaging Graph, 2017,66:100-114.DOI:10.1016/j.compmedimag.2017.11.001.

[14] Ghosal R , Maity A , Clark T , Longo S B. Variable selection in functional linear concurrent regression[J]. Applied Statistics, 2020, 69(3):565-587.Cite:arXiv:1904.08507.

[15] Tibshirani R. Regression shrinkage and selection via the LASSO[J]. Journal of the Royal Statistical Society, Series B, 1996, 58(1):267-288.DOI:10.1111/j.1467-9868.2011.00771.x.

[16] Yuan M, Lin Y. Model selection and estimation in regression with grouped variables[J]. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2006, 68:49-67.DOI:10.1111/j.1467-9868.2005.00532.x.

[17] Meie L, van de Geer S, Bühlmann P. The group LASSO for logistic regression[J]. Journal of the Royal Statistical Society Series B (Statistical Methodology), 2008, 70(1):53-71.DOI:10.1111/j.1467-9868.2007.00627.x.

[18] Breheny P, Huang J. Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection[J]. The Annals of Applied Statistics, 2011, 5(1):232-253.DOI：10.1214/10-AOAS388.

[19] Breheny P, Huang J. Group descent algorithms for nonconvex penalized linear and logistic regression models with grouped predictors[J]. Statistics and Computing, 2015,25(2):173-187.DOI:10.1007/s11222-013-9424-2.

[20] Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties[J]. Journal of the American Statistical Association,2001, 96:1348-1360.DOI:10.1198/016214501753382273.

[21] Zhang C. Nearly unbiased variable selection under minimax concave penalty[J]. The Annals of Statistics, 2010, 38:894-942.DOI:10.2307/25662264.

[22] Raghavan R, Ashour F S, Bailey R. A review of cutoffs for nutritional biomarkers[J]. Advances in Nutrition, 2016,7(1):112-120.DOI:10.3945/an.115.009951.