Jestr

Journal of Engineering Science and Technology Review 11 (2) (2018) 135 - 142

Research Article

JOURNAL OF
Engineering Science and Technology Review

www.jestr.org

Prognosis Evaluation of Ovarian Granulosa Cell Tumor Based on Co-forest Intelligence Model

Xin Liao[1,2], Xin Zheng[3], Juan Zou[1,2], Min Feng[1,2], Liang Sun[1,2], Yan Li[4] and Kaixuan Yang[1*]

[1] *Department of Pathology, West China Second University Hospital, Chengdu 610041, China*
[2] *Key Laboratory of Birth Defects and Related Disease of Women and Children, Ministry of Education, Sichuan University, Chengdu 610041, China*
[3] *College of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China*
[4] *Department of Immunology, Cleveland Clinic, Cleveland, Ohio 44195, United States*

Received 23 November 2017; Accepted 2 April 2018

Abstract

Ovarian granulosa cell tumor (GCT) has different recurrence periods, which dramatically decreases after the 5-year survival period. Prognosis evaluation has important clinical values and is a research hotspot. Prognosis evaluation methods include logistic regression, Chi-square analysis, and other traditional statistical methods; however, these techniques cannot solve problems, such as limited samples and ambiguous prognosis-related pathologic features, and have poor reliability and validity of assessment results. In this study, an artificial intelligence theory was introduced, and the prognosis evaluation of ovarian GCT based on co-forest intelligence model was proposed to find a method applicable to the pathological data of ovarian GCT with limited samples and ambiguous prognosis features. First, data preprocessing of ovarian GCT samples was performed. This procedure included deleting unqualified data and standardizing and normalizing data. Second, prognosis evaluation of ovarian GCT was accomplished by using co-forest intelligence algorithm. Finally, the validity of the proposed prognosis evaluation method was verified by 75 patients with ovarian GCT in the West China Second Hospital of Sichuan University. Results indicate that: (1) the accuracy of prognosis evaluation based on the feature set selected by Log-Rank test increases by 12.1% compared with that (4.1%) based on the direct use of standardized and normalized feature set, and (2) the co-forest algorithm can be used for the model analysis of small pathological datasets of ovarian GCT. Moreover, this method can be used to explore effective characteristics from the candidate feature dataset through automatic learning with prediction accuracy of up to 95.7%. This study reveals the reliability and effectiveness of the proposed prognosis evaluation method of ovarian GCT based on co-forest intelligence model. Conclusions are beneficial for clinicians to accurately understand the development laws of ovarian GCT, take the initiative to master the diagnosis and treatment, and increase the long-term survival rate of patients.

Keywords: Ovarian granulosa cell tumor, Prognosis evaluation, Log-Rank test, Co-forest intelligent algorithm

1. Introduction

Ovarian granulosa cell tumor (GCT) is a sex cord-stromal tumor with low grade malignancy (including adult GCT and juvenile GCT). Menstrual disorder in the reproductive age or irregular vaginal bleeding in menopause period, abdominal pain, pelvic mass and ascite, long-term recurrence, and significant reduction of five-year survival rate after recurrence are common clinical symptoms of ovarian GCT[1]. Therefore, prognosis evaluation of ovarian GCT is important for its diagnosis and treatment formulation by clinicians. A stable and reliable evaluation method is conducive to clinicians to take the initiative to master diagnosis and treatment and increase the long-term survival rate of patients. However, the existing prognosis evaluation methods of ovarian GCT mainly use traditional statistical approaches, such as logistic regression and Chi-square analysis. These methods determine the correlation between single factor and tumor recurrence[2]. Differences of relevant prognosis pathologic features remain unknown. Different literatures report significantly different outcomes, providing difficulty to clinicians reliable to find references for the diagnosis and treatment of this disease[3]. In addition, the long recurrence of ovarian GCT causes high loss ratio of follow-up, resulting in limited samples and further increasing difficulties against prognosis evaluation.

Recently, semi-supervised learning technology has broken the application bottleneck of modeling analysis based on small-sized dataset[4][5]. Semi-supervised learning technology can train the initial model based on few labeled ovarian GCT samples, predict the unlabeled ovarian GCT samples based on the automatic marking strategy of probability learning theory, and improve the generalization ability of the model learned and acquired from few labeled samples by using the effective information hidden in unlabeled data. These characteristics make artificial intelligence (AI) technology applicable to prognosis evaluation of ovarian GCT with limited samples and ambiguous relevant prognostic features. A series of associated features, including clinical and pathological features, is enlisted based on existing studies[3, 6-16].

*E-mail address: huaxipath@aliyun.com

doi:10.25103/jestr.112.19

Furthermore, the prognosis evaluation method of ovarian GCT based on co-forest algorithm and multiple factors was constructed to solve problems of limited ovarian GCT samples and ambiguous prognostic features.

2. State of the art

Abundant academic studies on prognosis evaluation of ovarian GCT have been reported at home and abroad, trying to find stable and reliable prognostic features of ovarian GCT and provide some references for postoperative treatment and therapeutic effect evaluation of tumors. Based on literature review, the recurrence period of ovarian GCT is in 5 years after the first visit[6]. In the study of Fox et al., more than 50% patients suffered recurrence in 2 years[3][6]. Schwartz et al. also reported that 76.3% patients suffered recurrence in 3 years[7][8]. To have patients who suffered recurrence after more than 10 years is also common[11][12]. Sommers once reported that sixpatients with ovarian GCT suffered recurrence after 20 years of operation[13]. The longest period of recurrence of ovarian GCT reaches 37 years[6]. Clinical features, such as recurrence of ovarian GCT, pelvic spread, and tumor involvement of extra ovarian organs, are believed as effective features of poor prognosis of ovarian GCT[6][14]. Various pathologic features of ovarian GCT are significantly correlated with clinical prognosis. However, different research conclusions still had many contradictions and disputes. Haba et al.[15] pointed out pathologic features with tumor well-differentiation. For example, follicular pattern of tumor cells and occurrence of Call-Exner body all promoted good tumor prognosis. Insular or diffuse pattern of tumor cells prompted poor differentiation of tumors and poor prognosis[15]. Pectasides et al.[16] believed that nuclear mitosis activity of tumor cells is related with the associated marker Ki-67 index, and expression levels of oncogene and anti-oncogene markers (e.g., P53, P16, and PTEN) are pathological features related with prognosis of ovarian GCT. However, no agreement on these research conclusions has been reached yet. Moreover, ovarian GCT is not a common ovarian tumor and has very limited clinical samples and difficult data availability (acquisition of one sample covers multiple programs, including collection of clinical data, pathologic image, and immunohistochemical staining).

The prognosis analysis modeling of ovarian GCT based on clinical and pathologic data shall prevent unqualified samples in pathologic dataset involved in iterative tuning of the model. Before the modeling analysis based on the feature dataset, the feature dataset has to be standardized and normalized. The correlation between pathologic features and clinical prognosis of ovarian GCT has not been determined completely. Therefore, this modeling requires that the applied intelligence algorithm shall be able to explore effective features from candidate feature dataset through automatic learning. In addition, with limited pathologic data of ovarian GCT, the intelligence algorithm shall be capable to establish the initial model by using few samples. Furthermore, the model can screen qualified samples for iterative tuning to improve the prediction performance of the model. Based on the above analysis, a prognosis evaluation method of ovarian GCT was proposed based on the co-forest intelligence model. First, a series of features, including clinical and pathologic features, was enlisted in the proposed model with reference to existing literatures and research results. Unqualified samples in the feature dataset were eliminated, and data standardization and normalization were performed. Subsequently, the co-forest intelligence algorithm that can explore effective features automatically was applied to prognosis evaluation of pathologic data of ovarian GCT.

The remainder of this study is organized as follows. Section 3 describes the research methodologies, including data preprocessing and co-forest intelligence algorithm for ovarian GCT prognosis evaluation. Section 4 constructs the GCT pathologic dataset based on 75 patients with ovarian GCT from April 2002 and February 2014 in West China Second Hospital of Sichuan University. Section 5 carries out the corresponding experiments and analyses based on above GCT pathologic dataset. Section 6 presents the conclusions.

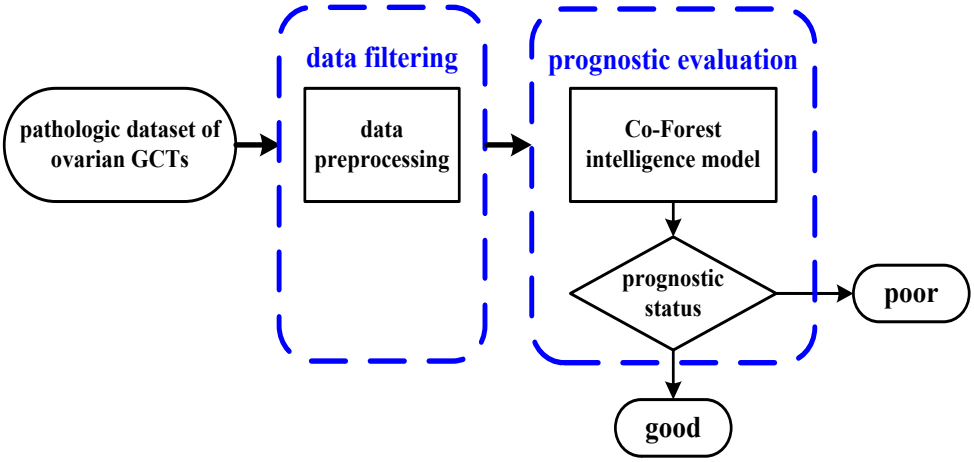

Fig. 1. Flowchart of the proposed prognosis evaluation method

3. Methodology

The flow chart of the proposed prognosis evaluation method of ovarian GCT based on the co-forest intelligence model is shown in Fig.1. First, pathologic data of ovarian GCT samples were preprocessed, including elimination of unqualified data and data standardization and normalization. Second, prognosis evaluation of ovarian GCT was accomplished by the co-forest intelligence algorithm. Structure of GCT pathologic data was determined with reference to previous literatures and research findings. A series of relevant features including clinical and pathological characteristics was enlisted in the proposed method.

3.1 Preprocessing of pathologic dataset of ovarian GCT

Original data are generally incomplete, redundant and fuzzy. The interference information in the original data may cause analysis bias[17]. Therefore, data preprocessing is needed before prognosis evaluation by using the co-forest intelligence model, including elimination of unqualified data according to expert rules and data standardization and normalization[18].

First, uniqueness, integrity (whether key attribute values in data records are clear and integral), validity (whether the value range of each attribute in data record is reasonable and mets the constraints), and consistency (whether unit of each attribute in data records is set uniform) should be observed. Inconsistent standards and data structure shall be avoided, and pathologic data of ovarian GCT shall be verified according to their medical significance. Data that fail to meet the above conditions shall be deleted.

Given that the feature dataset of ovarian GCT covers different types and dimensions of attributes, which may influence the modeling analysis results, data standardization and normalization are necessary to ensure that all features are at the same order of magnitudes and applicable for contrast analysis. Data standardization method is related with actual meaning and valuing mode of data and shall be judged according to expert rules. The data processing model will be interpreted in detail in Section 4.1. Zero-mean normalization method was adopted as follows:

$$N_k[i]=\frac{S_k[i]-M_k[i]}{\sqrt{V_k[i]}} \tag{1}$$

where $N_k[i]$ is the attribute i in the normalized sample k; and $S_k[i]$ is the attribute i in the sample k. $M_k[i]$ and $V_k[i]$ are mean and variance of attribute i in the sample k, respectively. The calculation formulas are shown as Eqs. (2) and (3).

$$M_k[i]=\frac{1}{m}\sum_{k=1}^{m}S_k[i] \tag{2}$$

$$V_k[i]=\frac{1}{m-1}\sum_{k=1}^{m}\left(S_k[i]-M_k[i]\right)^2 \tag{3}$$

where m is the sample size in the feature dataset of ovarian GCT.

3.2 Prognosis evaluation of ovarian GCT based on co-forest intelligence algorithm

Semi-supervised learning algorithm can train the initial model by using few labeled samples. During prediction of unlabeled samples, the model can screen unlabeled samples with high confidence coefficient for iterative tuning according to screening strategy, further improving the generalization ability of the model[19][20]. Co-training is an important branch in semi-supervised learning algorithm. Zhou et al. proposed the co-forest algorithm[21] based on the intelligent collaborative algorithm[4][5], which further used the collaborative performance of multiple basic models and can perform modeling analysis on small-sized dataset. Moreover, the co-forest algorithm is able to explore effective features from candidate feature dataset through automatic learning. In this study, the co-forest algorithm was applied for prognosis evaluation of ovarian GCT.

The co-forest model accomplishes the co-training by using six base classifiers. First, six independent sample subsets are acquired through Bootstrap resampling of labeled sample set and used to train six base classifiers. Next, unlabeled samples, which meet the requirements, are selected by combining classifiers (rest five base classifiers) as the supplementary sample set for iterative tuning of the model. The iterative training of the co-forest intelligence model is shown in Fig.2. Specific steps are introduced as follows.

Step 1) Six independent training sample subsets (L_1, L_2, L_3, L_4, L_5, and L_6) are constructed through Bootstrap resampling[22] from labeled sample set. They are used to train base classifiers (random stress[23]) (bc_1, bc_2, bc_3, bc_4, bc_5, and bc_6), which can explore effective features automatically from the original dataset.

Step 2) Implementing co-training of six base classifiers. The unlabeled samples that shall be added in base classifier bc_i for next iterative training are determined by voting of the combining classifier HC_i . Next, the newly constructed sample set is used to re-train base classifiers.

First, classification errors e_i (suppose it is the i-th iteration at present) of labeled sample set by the combining classifier HC_i (combination of five base classifiers except the base classifier bc_i) are recorded. If e_i meets Eq. (4), samples with high confidence coefficient, which meet the conditions, are selectedas the extended training set.

$$\begin{cases} e_{i,t} < e_{i,t-1} \\ W_{i,t} < W_{i,t-1} \end{cases} \tag{4}$$

where the initial value of classification error (e_0) can be set as 0.5. During optimization of base classifiers, extended sample set is only selected when the performance of the combining classifier is improved. Specifically, data in unlabelled sample set are added into the candidate extended sample set. When the weight sum of all added unlabeled samples is higher than the threshold, adding is stopped. Next, extended samples are screened from the candidate extended sample set according to the confidence coefficient. Single candidate sample in the candidate sample set, which has lower confidence coefficient than the threshold shall be deleted. Then, candidate sample set is formed by screening according to threshold of single confidence coefficient and subsequently judged by Eq. (5). If the sample meets the conditions, it is used as the extended sample set. Otherwise, the sampleis deleted.

$$W_{i,t} < \frac{e_{i,t-1} * W_{i,t-1}}{e_{i,t}} \tag{5}$$

where $W_{i,t}$ is the weight of the sample set at the i-th iteration. $W_{i,t}$ is calculated as follows.

The weight $W_{i,t,j}$ is the predicted confidence coefficient of sample x_j of the $n-1$ classifier except for i-th classifier at the t-th iteration.

According to the above method, training sample sets of base classifiers (bc_1, bc_2, bc_3, bc_4, bc_5, and bc_6) are extended by using the combining classifier(HC_1, HC_2, HC_3,

HC_4, HC_5, and HC_6). In this way, the co-training of six base classifiers is accomplished.

Step 3) Determining whether supplementary samples added to six base classifiers is judged one by one. If yes, the supplementary samples are integrated with current sample set of the base classifier to re-train the base classifier and update the state of the flag bit.

Step 4) The updating flag bit of six base classifiers is checked one by one. If none is updated, training of co-forest intelligence model is stopped. Otherwise, step 2 is performed, and the co-training is continued.

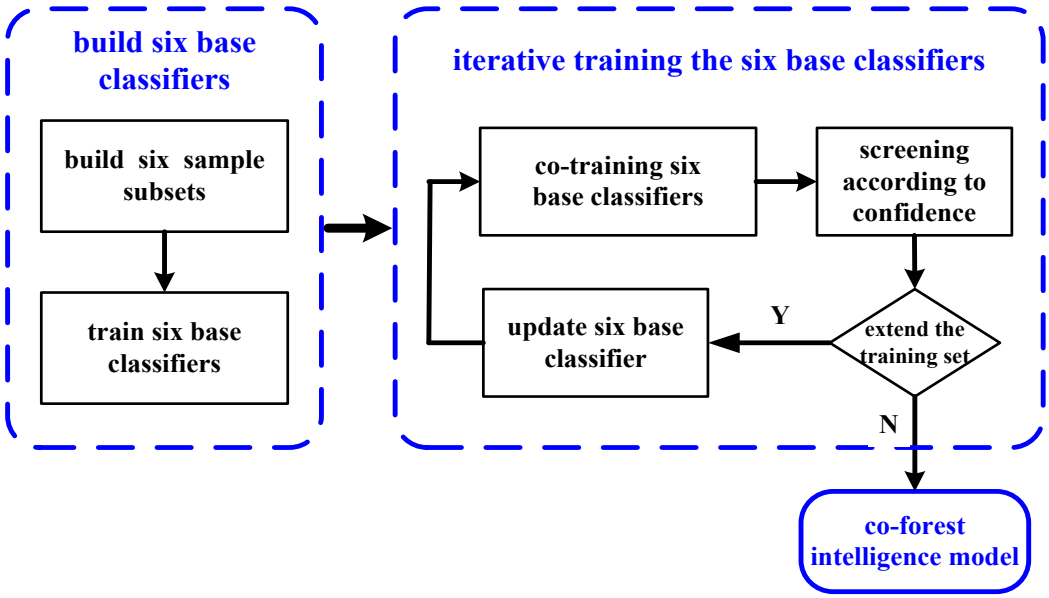

Fig. 2. Training of the co-forest intelligence model

3.3 Construction of pathologic dataset of ovarian GCT

Patients with ovarian GCT from April 2002 to February 2014 diagnosed and hospitalized in the West China 2nd University Hospital of Sichuan University were selected in the research based on the following rules.

(1) Diagnosis of ovarian GCT was reviewed and confirmed by senior pathologists.

(2) Complete clinical data from the first visit to treatment period.

(3) Follow-up visit≥3 years.

Finally, 75 patients with ovarian GCT were investigated in this experiment, including 17 patients suffering recurrence of tumor in the follow-up visit. The recurrence period ranges between 6 and 52 months, 32 months in average. Among them, three patients died of recurrence.

Clinical data of all patients were reviewed, and different clinical characteristic features were summarized, including age, modus operandi, clinical stage of tumor, and postoperative chemotherapy. In the pathologic dataset of ovarian GCT, patients aged from 14 to 80, and the age of median onset was 47 years old. All patients were treated by operations. Among them, 42 patients (56%) had primary operation and adopted uterus + bilateral adnexectomy +/− lymph node excision, 24 patients (32%) had adnexectomy of the affected side or tumorectomy, and 9 patients (12%) had tumor reductive surgery. After the operation, 49 patients (52%) were determined as stage I, 20 patients (40%) at stage II, and 6 patients (8%) at stage III. In addition, 48 patients (64%) received radiotherapy/chemotherapy, and another 27 patients (36%) had not received radiotherapy/chemotherapy.

Pathologic data and sections of all patients are reviewed by senior attending doctors. The tumor diameter ranges between 2.5 and 14cm, 5.8cm in average. Specifically, six patients (8%) had spontaneous tumor rupture. Tumor patterns under a microscope are mainly follicular pattern (Fig.3), insular pattern (Fig.4), trabecular pattern (Fig.5), and diffuse/sarcoma pattern (Fig.6), accompanied with few combined patterns (two or more patterns in the above four patterns). Twenty six patients (34.7%) presented tumor hemorrhage and necrosis. Call-Exner body (Fig.7) was observed in 48 patients (64%) (Fig.7), and tumor luteinization was detected in 23 patients (30.7%) (Fig.8). The nuclear mitosis phase of tumor counted 1-21/10HPF, 7/10HPF in average. In immunohistochemical test, 52 patients (69.3%) were PTEN positive, 62 patients (82.7%) were p16 positive, and 43 patients (57.3%) were p53 positive (positive cells>50%[6]).

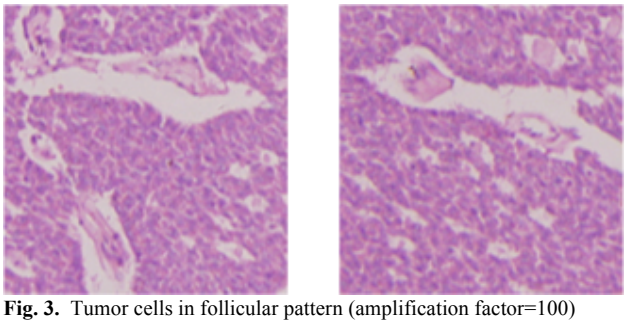

Fig. 3. Tumor cells in follicular pattern (amplification factor=100)

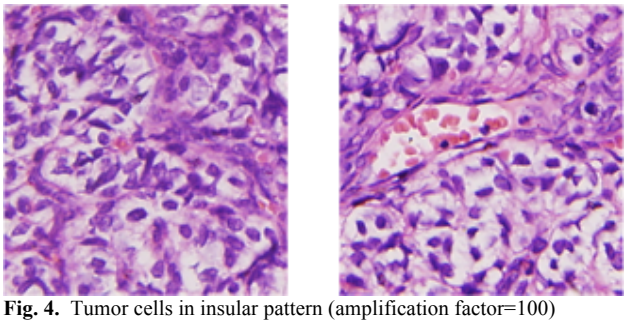

Fig. 4. Tumor cells in insular pattern (amplification factor=100)

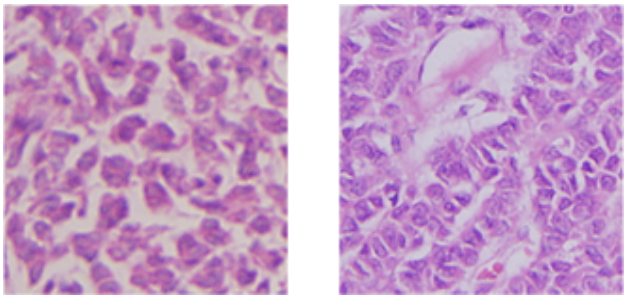

Fig. 5. Tumor cells in trabecular pattern (amplification factor=100)

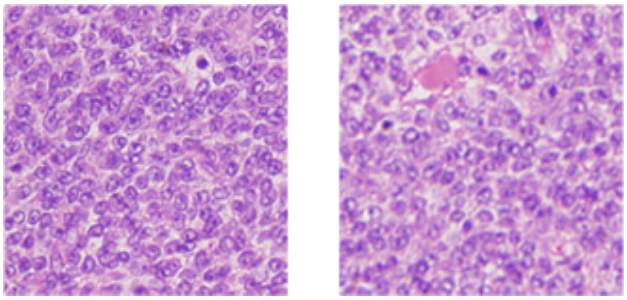

Fig. 6. Tumor cells in diffuse/sarcoma pattern (amplification factor=100)

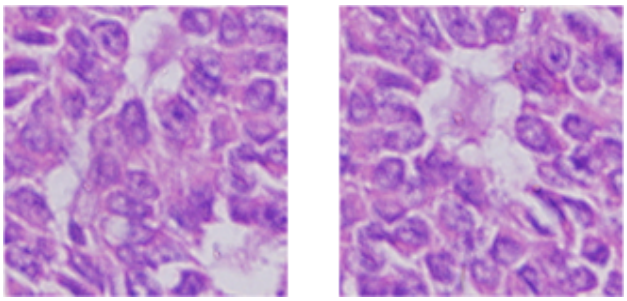

Fig. 7. Tumor cells in Call-Exner body pattern (amplification factor=400)

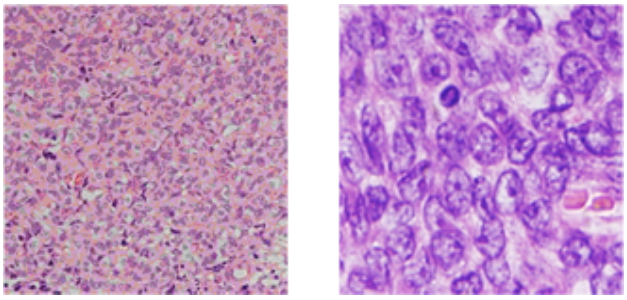

Fig. 8. Luteinization and nuclear mitosis of tumor cells (amplification factor=400)

Patients were divided into the recurrence group and the non recurrence group. Pathological and clinical factors related with tumor recurrence were analyzed preliminarily by Log-Rank test[24]. Clinical factors related with recurrence included clinical stage of tumor and postoperative chemotherapy ($p<0.05$). Pathological factors included spontaneous tumor rupture, tumor cell pattern (insular or diffuse patterns), nuclear mitosis number of tumor cells, and positive rates of p53 and Ki-67 index ($p<0.05$).

4 Simulation result analysis

4.1 Data preprocessing experiment and analysis

Pathologic sample set of ovarian GCT is preprocessed, including standardization and normalization. Preprocessing rules for specific data in samples are shown in Table 1. Different preprocessing rules are described as follows.

(1) Rule I: binary data term is 1 if it has correspondence. Otherwise, it values 0.

(2) Rule II: multiple-valued data term with determined value is discretized according to regulated proportion.

(3) Rule III: multiple-valued data term without determined value is truncated first according to the upper limit set by expert rules and then discretized.

Table. 1. Data preprocessing rules of ovarian GCT

Index	Rules
clinical stage of tumor	II
postoperative chemotherapy	I
Call-Exner body	I
number of nuclear mitosis	III
cell atypism	I
haemorrhage and necrosis	I
follicular pattern	I
insular pattern	I
trabecular pattern	I
ribbon pattern	I
diffuse pattern	I
luteinization of tumor cells	I
Ki-67 expression	II
PTEN	II
EGFR	II
P53	II
prognostic status	I

Pathological samples of ovarian GCT include 17 clinical/pathological features and 1 prognosis status. Some preprocessed pathological data samples of ovarian GCT are listed in Table 2, including the original data and preprocessed (standardized and normalized) data.

In Table 2, attribute values of all preprocessed data meet the requirements of standardization and normalization.

4.2 Experimental analysis on prognosis evaluation of ovarian GCT

For ovarian GCT samples, a series of relevant features including clinical and pathological features was enlisted with reference to previous literature and research results. On this basis, two feature sets were constructed, including the following.

(1) Feature set (M1) after standardization and normalization of all features in Table 1 except the prognosis status was established.

(2) Based on M1, the feature set (M2) of factors, which have significantly statistical ($p<0.05$) correlations with recurrence according to preliminary Log-Rank test, was constructed. It covers clinical stage of tumor, postoperative chemotherapy, nuclear mitosis, spontaneous tumor rupture, positive rate of immunohistochemical markers (p53 and Ki-67), and pattern of tumor cells (follicular and diffuse patterns).

For M1 and M2, the proposed co-forest intelligence model was applied for the experiment of prognosis prediction. The results were compared with the decision tree C4.5[25] and support vector machine (SVM) model [26]. The three-fold cross validation method was applied in the experiment. The receiver operator characteristic curve (ROC) of prognosis evaluation based on co-forest intelligence model, decision tree C4.5, and SVM model based on M1 are shown in Fig.9. The ROC curves of prognosis evaluation based on co-forest intelligence model, decision tree C4.5,

and SVM model based on M2 are shown in Fig.10. The ROC curves of prognosis evaluation based on co-forest intelligence model, decision tree C4.5, and SVM model based on M1 and M2 are shown in Figs.11–13. Prediction performance statistics of the above three models based on different feature sets are presented in Table 3.

Table. 2. Preprocessing results of some ovarian GCT data

Index	Sample 1		Sample 2	
	before	after	before	after
clinical stage of tumor	stage II	0.5	stage I	0
postoperative chemotherapy	none	0	Exist	1
call-Exner body	exist	1	Exist	1
number of nuclear mitosis	1	0.2	3	0.6
cell atypism	none	0	None	0
haemorrhage and necrosis	exist	1	None	0
follicular pattern	none	0	Exist	1
insular pattern	exist	1	Exist	0
trabecular pattern	exist	1	None	0
ribbon pattern	none	0	None	0
diffuse pattern	exist	1	None	0
luteinization of tumor cells	none	0	None	0
Ki-67 expression	50%	0.5	20%	0.2
PTEN	focal positive	0.33	focal positive	0.33
EGFR	negative	0	Negative	0
P53	negative	0	focal positive	0.33
prognostic status	favorable	1	unfavorable	0

Figs.9 and 10 show that the proposed prognosis evaluation method of ovarian GCT based on the co-forest intelligence model is superior decision tree C4.5 and SVM model in terms of prognosis prediction based on either M1 or M2. Figs.11–13 show that the prognosis prediction accuracies of the co-forest intelligence model, decision tree C4.5, and SVM model based on M2 are significantly higher than those based on M1. These experimental results prove the validity of preliminary feature set screening by Log-Rank test. Table 3 shows that the area under the ROC curves (AUCs) of the co-forest intelligence model based on M1 and M2 (0.916 and 0.958, respectively) are far larger than those of the decision tree C4.5 (0.741 and 0.862) and SVM model (0.713 and 0.798).

According to the above results, the proposed prognosis evaluation of ovarian GCT based on co-forest intelligence model has significantly higher validity than those of decision tree C4.5 and the SVM model. Furthermore, prognosis prediction accuracies of the co-forest intelligence model, decision tree C4.5, and SVM model based on M2 are higher than those based on M1, proving validity of Log-Rank test in selection of the original feature set. The proposed method overcomes problems of limited pathological samples and difficult determination of prognosis-relevant factors in prognosis evaluation of ovarian GCT. Furthermore, the method achieves satisfying prediction performance and has high practical value in prognosis evaluation. This study is conducive to clinicians to optimize the treatment scheme and realize individual precision treatment based on comprehensive evaluation of patients' conditions, thus guaranteeing the long-term survival rate and survival quality of patients.

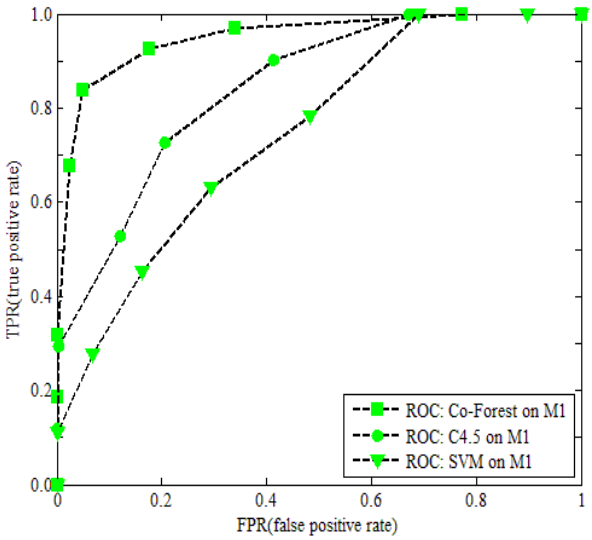

Fig. 9. ROC curves of evaluation models based on M1

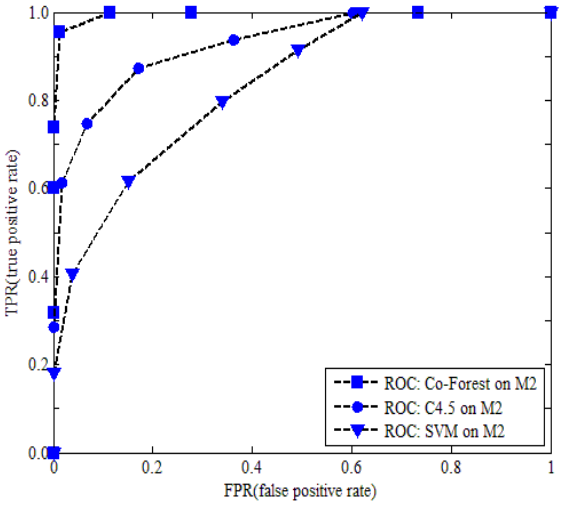

Fig. 10. ROC curves of evaluation models based on M2

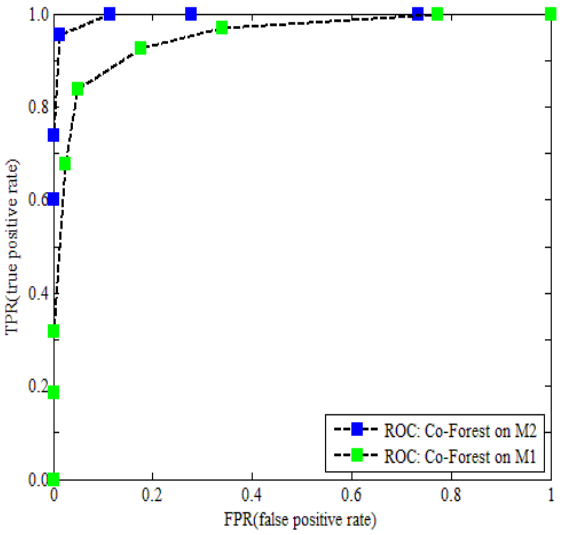

Fig. 11. ROC curves of the co-forest intelligence model based on M1 and M2

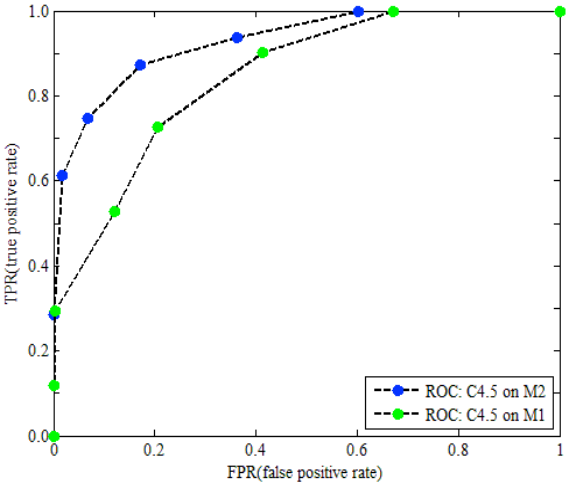

Fig. 12. ROC curves of the decision tree C4.5 based on M1 and M2

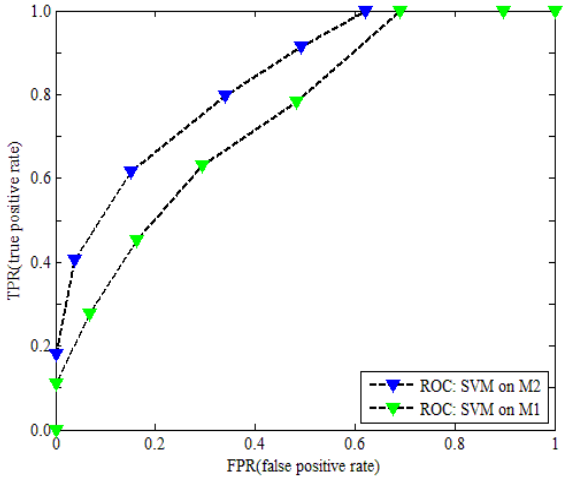

Fig. 13. ROC curves of the SVM model based on M1 and M2

Table. 3. Performances of different prognosis evaluation methods of ovarian GCT

Algorithm Model	AUC	
	M 1	M 2
co-forest model	0.916	0.957
decision tree C4.5 model	0.741	0.862
support vector machine	0.713	0.798

5. Conclusions

Ovarian GCT has limited samples and significant different periods of recurrence, resulting in many difficulties of prognosis evaluation. In this study, AI theory and machine learning technology are introduced into prognosis evaluation of tumors, and a prognosis evaluation method of ovarian GCT based on the co-forest intelligence model is proposed. Some conclusions can be drawn according to experimental results.

(1) The prognosis evaluation of ovarian GCT based on M1, which is standardized and normalized, is poorer than that based on M2, which is selected by Log-Rank test. Currently, pathological features and clinical features related with prognosis of ovarian GCT have not been determined completely. Therefore, M1 must have some invalid and even interference features. Log-Rank test can eliminate some interference features, thus improving the prediction accuracy of the model.

(2) The co-forest intelligence model can make modeling analysis on small-sized dataset and explore effective features from candidate feature dataset through automatic learning. It overcomes some shortcomings of ovarian GCT prognosis (i.e., incomplete determination of relevant pathological and clinical features) and achieves satisfying prognosis prediction results. The AUCs of co-forest intelligence model based on M1 and M2 (0.916 and 0.958, respectively) are far larger than those of the decision tree C4.5 (0.741 and 0.862) and SVM model (0.713 and 0.798).

The proposed prognosis evaluation method of ovarian GCT based on co-forest intelligence model not only overcomes limited sample data and ambiguous prognostic features but also achieves good prediction results. It has high practical value in prognosis evaluation. Research conclusions are conducive to break bottlenecks against prognosis evaluation of ovarian GCT and can help clinicians master development laws of ovarian GCT, take the initiative in diagnosis and treatment, and increase long-term survival rates of patients. However, further improvements are still needed. Future studies shall further collect ovarian GCT samples to increase the generalization of the prognosis prediction model.

Acknowledgements

The authors are grateful for the support provided by the Program of Key Laboratory Open Fund in Sichuan Province (Grant No. 2017LF3008) and the Important Special Fund for Applied R & D in Guangdong Province (Grant No. 2015BD10131002).

References

1. Färkkilä, A., Haltia, U. M., Tapper, J., et al., "Pathogenesis and treatment of adult-type granulosa cell tumor of the ovary". *Annals of Medicine*, 49(5), 2017, pp.435-447.
2. Nosov, V., Silva, I,, Tavassoli, F., et al., "Predictors of recurrence of ovarian granulosa cell tumors". *International Journal of Gynecological Cancer*, 19(4), 2009, pp.628-633.
3. Klemi, P. J., Joensuu, H., Salmi, T., "Prognostic value of flow cytometric DNA content analysis in granulosa cell tumor of the ovary". *Cancer,*65(5), 1990, pp.1189-1193.
4. Zhou, Z. H., Li, M., "Semisupervised regression with cotraining-style algorithms". *IEEE Transactions on Knowledge & Data Engineering*, 19(11), 2007, pp.1479-1493.
5. Raahemi, B., Zhong, W., Liu, J., "Exploiting unlabeled data to improve peer-to-peer traffic classification using incremental tri-training method". *Peer-to-Peer Networking and Applications*, 2(2), 2009, pp.87-97.
6. Khosla, D., Dimri, K., Pandey, AK., et al., "Ovarian granulose cell tumor: clinical features, treatment, outcomes, and prognostic factors". *North American Journal of Medical Sciences*, 6(3), 2014, pp.133–138.
7. Sehouli, J., Drescher, F. S., Mustea, A., et al., "Granulosa cell tumor of the ovary: 10 years follow-up data of 65 patients". *Anticancer Research*, 24(2C), 2004, pp.1223-1229.

8. Liao, X., Feng. M., Wang. H., "Pathologic features and prognostic factors of ovarian granulosa cell tumor". *Journal of Sichuan University (Medical science edition)*, 44(3), 2013, pp.419-423.
9. Fox, H., Agrawal, K., Langley, F. A., "A clinicopathologic study of 92 cases of granulosa cell tumor of the ovary with special reference to the factors influencing prognosis".*Cancer*, 35(1), 1975, pp.231-241.
10. Schwartz, P. E., Smith, J. P., "Treatment of ovarian stromal tumors". *American Journal of Obstetrics & Gynecology*, 125(3), 1976, pp.402-411.
11. Cheong, M. L., Shen, J., Huang, S. H., "Long-term survival in a patient with an advanced ovarian juvenile granulosa cell tumor with para-aortic lymph node metastasis". *Taiwanese Journal of Obstetrics & Gynecology,* 55(6), 2016, pp.907-909.
12. Majdoul, S., Tawfiq, N., Bourhaleb, Z., "Recurrence occurring ten years after the initial diagnosis of granulosa cell tumour of the ovary: about two cases and review of the literature". *Pan African Medical Journal*, 51(4), 2016, pp.25-30.
13. Sommers, S. C., Gates, O. Goodof II, "Late recurrence of granulosa cell tumors: report of two cases". *Obstetrics & Gynecology*, 6(4), 1955, pp.395-398.
14. Seagle, B. L., Ann, P., Butler, S., Shahabi, S., "Ovarian granulosa cell tumor: anational cancer database study". *Gynecologic Oncology*, 146(2), 2017, pp.285-291.
15. Haba, R., Miki, H., Kobayashi, S., et al., "Combined analysis of flow cytometry and morphometry of ovarian granulosa cell tumor". *Cancer*, 72(11), 2015, pp.3258-3262.
16. Pectasides, D., Pectasides, E. A., "Granulosa cell tumor of the ovary".*Cancer Treatment Reviews*, 34(1), 2008, pp.1-12.
17. H,M.W., Lin, W.C., Chen,C.W., et al., "Data preprocessing issues for incomplete medical datasets". *Expert Systems*, 33(5), 2016, pp.432-438.
18. Haustein, S., "Grand challenges in altmetrics: heterogeneity, data quality and dependencies". *Scientometrics*, 108(1), 2016, pp.413-423.
19. Zhang, K., Lan, L., Kwok, J. T., et al., "Scaling up graph-based semisupervised learning via prototype vector machines". *IEEE Transactions on Neural Networks & Learning Systems*, 26(3), 2017, pp.444-457.
20. Yeung, D. Y., Chang, H., Dai, G., "A scalable kernel-based semisupervised metric learning algorithm with out-of-sample generalization ability". *Neural Computation*, 20(11), 2008, pp.2839-2861.
21. Li, M., Zhou, Z. H., "Improve computer-aided diagnosis with machine learning techniques using undiagnosed samples". *IEEE Transactions on Systems, Man, and Cybernetics*, 29(3), 2007, pp.1088-1098.
22. Robinson, A., "Randomization, bootstrap and monte-carlo methods in biology". *Journal of the Royal Statistical Society*, 170(3), 2010, pp.856-859.
23. Yang, R. M., Zhang, G. L., Liu, F., et al., "Comparison of boosted regression tree and random forest models for mapping topsoil organic carbon concentration in an alpine ecosystem". *Ecological Indicators*, 60(2), 2016, pp.870-878.
24. Koletsi, D., Pandis, N., "Survival analysis, part 2: kaplan-meier method and the log-rank test". *American Journal of Orthodontics & Dentofacial Orthopedics*, 152(4), 2017, pp.569-571 .
25. Zouggar, S.T., Adla, A., "Proposal for measuring quality of decision trees partition". *International Journal of Decision Support System Technology*, 9(4), 2017, pp.16-36.
26. Wu, J., Yang, H., "Linear regression-based efficient svm learning for large-scale classification". *IEEE Transactions on Neural Networks & Learning Systems*, 26(10), 2017, pp.2357-2369.