Mining a Small Medical Data Set by Integrating the Decision Tree and *t*-test

Ming-Yang Chang
Department of Obstetrics and Gynecology, Chang Gung Memorial Hospital, Taipei, Taiwan 25137, R.O.C.
Email: mychang1126@yahoo.com.tw

Chien-Chou Shih[2],Ding-An Chiang[1] and Chun-Chi Chen[1 *]
[1] Department of Computer Science & Information Engineering, Tamkang University, Tamsui, Taipei County, Taiwan 25137, R.O.C.
[2] Department of Information & Communication, Tamkang University, Tamsui, Taipei County, Taiwan 25137, R.O.C.
Email: ccs@mail.tku.edu.tw, chiang@cs.tku.edu.tw, maurice.chen.tw@gmail.com

***Abstract*—Although several researchers have used statistical methods to prove that aspiration followed by the injection of 95% ethanol left in situ (retention) is an effective treatment for ovarian endometriomas, very few discuss the different conditions that could generate different recovery rates for the patients. Therefore, this study adopts the statistical method and decision tree techniques together to analyze the postoperative status of ovarian endometriosis patients under different conditions. Since our collected data set is small, containing only 212 records, we use all of these data as the training data. Therefore, instead of using a resultant tree to generate rules directly, we use the value of each node as a cut point to generate all possible rules from the tree first. Then, using *t*-test, we verify the rules to discover some useful description rules after all possible rules from the tree have been generated. Experimental results show that our approach can find some new interesting knowledge about recurrent ovarian endometriomas under different conditions.**

***Index Terms*—Data mining, Decision tree, *t*-test, p-value, Ovarian endometriomas**

I. INTRODUCTION

The use of classification algorithms in the medical domains has increasingly been the object of study in recent years [1-3]. Although many conventional treatments are available in clinical medicine for patients suffering from endometriosis, it is a common disease among women of reproductive age with a high recurrence rate, regardless of the treatment type [4]. In recent years, ultrasound-guided aspiration combined with drug therapy has become a new alternative for patients since the ultrasound-guided aspiration of ovarian endometriomas was proposed in 1991 [5]. Therefore, to reduce the high recurrence rate, some medical treatments have combined ultrasound-guided aspiration with tetracycline [6], methotrexate [7], recombinant interleukin-2 [8], or ethanol [9-11].

In recent years, several researchers and our team have used statistical methods to prove that aspiration followed by the injection of 95% ethanol left in situ (retention) is an effective treatment for ovarian endometriomas [11]. They always divide all patients into two groups group 1 (ethanol irrigation) and group 2 (ethanol retention) regardless of other conditions, such as the size of the cyst. In endometriomas-like dataset, the analysis of the influence of treatment effectiveness use statistical and *t*-test cannot integration data mining in related research [12-16] such as *t*-test [12], Logistic regression [15, 17], Decision trees [18-20], and SVM [15, 17]. However, different conditions could generate different recovery rates for the patients, and very few researchers discuss this situation. Therefore, our aim is to investigate recurrent ovarian endometriomas under different conditions. We adopt the statistical method and decision tree techniques together to analyze the postoperative status of ovarian endometriosis patients.

The use of machine learning algorithms for the building of predictive and descriptive data mining models has become widely accepted in medical applications. Various models including Decision trees, Decision rules, Logistic regression, Artificial neural networks, and SVMs have been tested in a wide variety of clinical and medical applications [21]. In order to resolve the endometriomas problem, we should identify and treat the cause of the problem correctly. But such correct diagnosis and treatment require the patients to have extensive live, womb, uterocervical canal, laparoscopy, and others [22]. Computationally, there have been attempts from the endometriomas medical domain to analyze various treatment factors to predict the success of therapy. Previous studies that have utilized *t*-test for endometrial related research have only used this technology to determine the effectiveness of treatment [23]. Based on the decision-tree analysis, the optimal rule to detect the ultrasound characteristics of endometriomas in pre- and postmenopausal patients and to develop rules that characterize endometriomas [20]. For ovarian tumors and pregnancies of unknown location on medical

* Corresponding author.

doi:10.4304/jsw.6.12.2515-2520

decision making (prediction), mathematical algorithms are applied to data sets in order to obtain a model [15, 16]. Kinkel [24] et al. demonstrated in a meta-analysis of indeterminate masses in sonography the superiority of MRT over CT and Doppler sonography in predicting malignancy. Fewer models have been hitherto developed, even though the reliable identification of borderline and ovarian endometriomas would be a good step forward for clinical practice.

The decision tree in classification algorithms has been applied to categorical attributes and numeric attributes in different domains [25]. Since medical data always contain numeric attributes and handling them is a critical task in inductive learning, this task has already been embedded within the decision tree algorithm. Therefore, we adopt the decision tree technique to analyze postoperative status of ovarian endometriosis patients in this study. The decision tree is built by performing a heuristic-based local search to select the best test attribute as the root of the decision tree. So, decision tree creates a branch for each value of that appearing in the training data. Then, the same procedure is operated on each branch to induce the remaining levels of the decision tree until all examples in a leaf belong to the same class. Whereas the tree is represented by a set of rules, each branch of the decision tree is only represented by a rule. However, since each branch of the decision tree is only represented by a rule and the resultant rules of the tree are local, some useful description rules cannot be found by the tree-generation algorithm. Therefore, instead of using a resultant tree to generate classification rules directly, we use the value of each node as a cut point to generate all possible rules from the tree. We can filter out useless rules by using the method of setting minimum recovery rates. However, in the study case, our collected data set is small, containing only 212 records, and we use all of these data as the training data. So because there are different conditions in a possible rule that could generate different recovery rates for the patients and very few researchers discuss this situation, our aim is to investigate the recurrent ovarian endometriomas under the different conditions in possible rule. Under this mining goal, we use *t*-test to verify rules to discover some useful description rules after all possible rules from the tree are generated.

The models were built in collaboration with gynecologists, and resulted in accurate predictions. The ovarian endometriomas models were based on multi-center data and have successfully passed prospective internal and external evaluations. The models for pregnancy of unknown location were based on single center data and have passed the first internal evaluation. However, a large multi-center study is ongoing, aiming for a thorough validation and for the development of new models for pregnancies of unknown location.

II. Materials and Background Knowledge

A. Patients

This study was approved by the Institutional Review Board of our hospital. It was a retrospective review of 212 consecutive patients treated at the outpatient gynecological department of Chang Gung Memorial Hospital, Taipei, Taiwan from July 1994 to July 2008. All patients had undergone previous surgical treatment for ovarian endometriomas and were being seen because of a recurrence. Recurrence was defined as when one or more persistent pelvic cysts greater than 3.0 cm were detected in two consecutive ultrasonographic examinations. Transvaginal ultrasound-guided cyst aspiration and ethanol injection were done on an outpatient basis. Patients randomly received ethanol instillation at 0 min (ethanol injected, then immediately removed), 3–9 min, 10 min, or retention. The procedure was in accordance with that reported previously [8, 23] .

The medical records include basic information of patients, treatment-related information, and clinical examination data obtained from the first visit and three-month, six-month to one year follows-up. The medical data set has 57 attributes, such as age, number of previous pregnancies, degree of menstrual cycle pains, urge technology to help the patient get pregnant, size of the cyst, CA125 blood test value, types of surgery, ethanol irrigation duration. Usually this raw data contained lots of null values (missing information), which might have affected the accuracy of the study. In clinical practice data collection is difficult because the value of attention to patient privacy; so data and collection methods are limited. We required a preprocessing procedure to prepare the valid data for the analysis.

B. Decision Trees

A decision tree is built by selecting the best test attribute as the root of the decision tree. Then, the same procedure is operated on each branch to induce the remaining levels of the decision tree until all examples in a leaf belong to the same class. Decision trees can be categorized by data processing functions into classification trees and regression trees. The common decision tree algorithms are compared in Table I. A classification tree is applied to discrete variables, while a regression tree is applied to quantitative variables. The regression tree was first brought up by Breiman [26] in the introduction of Classification and Regression Trees

TABLE I.
COMPARISON OF DECISION TREE ALGORITHMS

Name	Data attribute	tree type	Partitioning rule	Pruning rule
ID3	Discrete	Classification tree	Information Gain	No Pruning
C4.5	Numeric	Regression tree	Information Gain	Predicted error rate
CHAID	Class	Classification tree	Chi-Square test	No Pruning
CART	Discrete and Numeric	Classification and regression tree	Gini index	Entire error rate

(CART). Usually, in CART analysis, data is categorized into quantitative and discrete data. Quantitative data can be applied to prediction, while discrete data can be applied to classification. In other words, CART analysis is able to simultaneously process both quantitative and discrete data. In the article, we use CART trees in the software of DB2 Intelligent Miner for Data, version 8.1 (IBM Corp., New York), to analyze the collected data to generate the resultant tree.

Ⅲ. MINING USING DECISION TREE AND RESULTS

In the article, we use the CART function in IBM DB2 Intelligent Miner for Data, version 8.1, to analyze the collected data to generate the resultant tree. The first step in reducing the dimensionality is to use PCA algorithm by using the dataset, and decision tree is used for the second feature and rule extraction. After that in each iteration of the loop (step 1) confirmation attributes is selected into the output set (step 3). The selected classification attributes included Cyst_size, CA_125, and BMI. Although the data set has many attributes, as shown in figure 1, only four attributes are selected by the system to build the decision tree. These four attributes are Cyst_size, CA_125, BMI, and Recovery (Recovery means no operation within six months and pregnant within six months or cyst size less than 3 cm in the sixth month). Therefore, when the patient has recovered, the value of this attribute is set to 1; otherwise, it is set to 0. Moreover, Cyst_size, CA_125, and BMI are the size of cyst, blood test value, and body mass index before operation, respectively. The resultant rules generated from the decision tree are shown as follows:

(1) If Cyst_size ≥ 4.25 Then Recovery = 0 (recovery rate =32%)

(2) If Cyst_size < 4.25 and CA_125 ≥ 146.81Then Recovery = 0 (recovery rate =0%)

(3) If Cyst_size < 4.25 and CA_125 < 146.81and BMI < 22.7516 Then Recovery = 1 (recovery rate =80.77%)

(4) If Cyst_size < 4.25 and CA_125 < 146.81and BMI ≥ 22.7516 Then Recovery = 0 (recovery rate =25%)

The CART is a binary tree, as shown in figure 1, there is a significant difference between two branches of a node when the pruning process is performed automatically by the mining tool. Therefore, we can interpret the corresponding rules from the tree directly. However, to strengthen our conclusion, we still divided all the patients are into two different groups according to the corresponding rules obtained from the tree. The significance level is 0.05. The recovery rate of these 212 patients is 43.4%.

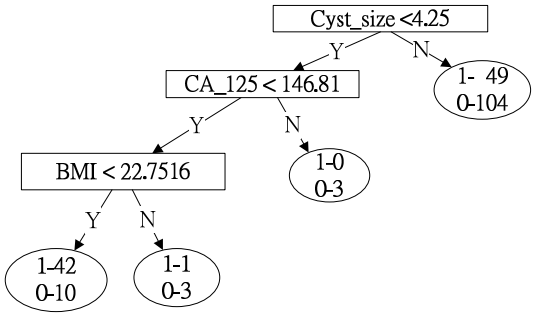

Figure 1. The resultant decision tree.

As shown by test 1 in Table II, according to the rule "Cyst_size ≥ 4.25 ==> Recovery = 0" all patients are divided into two groups by the cut point of the Cyst_size attribute. Group 1 contains patients whose cyst size is less than 4.25, while cyst size of patients in group 2 are not less than 4.25. From test 1 in the table, we conclude that when all patients are divided into two groups: group 1 (Cyst_size <4.25) and group 2 (Cyst_size ≥ 4.25). The recovery rate of group 1 (43/59, 72.88%) is significantly (p = 1.47E-08) greater than that of group 2 (49/153, 32%). In other words, the recovery rate (or recurrence rate) is affected by the cyst size regardless of the treatment type. Moreover, from test 2 and 3 in Table Ⅱ, we also conclude that the recovery rate of group 1 is significantly better than that of group 2.

Since each branch of the decision tree is only represented by a rule, and the resultant rules of the tree are local [27], some useful description rules cannot be found by the tree-generation algorithm. For example, the rule "Cyst_size < 4.25 and CA_125 ≥ 146.81 ==> Recovery = 0" should be ignored in the discussion because only three of the patients are related to this rule and the over-fit problem. Consequently, the question is "for those patients whose Cyst_size < 4.25, whether the BMI value will affect the recovery rate." Since our data size is very small and there are only three cut points in this example, to overcome the above problems, we could use these three cut points to generate all possible rules from the tree. We discuss this phenomenon in the

TABLE II.
COMPARISON OF DECISION TREE ALGORITHMS

test	Group 1	Group 2	Group 1 Recovery-rate (y/total, rate)	Group 2 Recovery-rate (y/total, rate)	*p*-value
1	Cyst_size <4.25	Cyst_size≥4.25	43/59. 72.88%	49/153, 32%	1.47E-08
2	Cyst_size <4.25 and CA_125 <146.81	Cyst_size <4.25 and CA_125 ≥146.81	41/53, 77.36%	0/3, 0%	0.001354
3	Cyst_size <4.25 and CA_125 <146.81 and BMI< 22.7516	Cyst_size <4.25 and CA_125 <146.81 and BMI≥22.7516	36/43, 83.72%	1/4, 25%	0.002647

following section.

Ⅳ. Interpreting Mining Results Regardless to Treatment Type

From the tree in figure 1, we obtain the cut points of Cyst_size, CA_125, and BMI numeric attributes as 4.25, 146.81, and 22.752, respectively. We can use these three values to generate all possible rules from the tree. Because the data set only contains 212 records, it is small. So, we used all of these data, as in training data. However, there are different conditions for generating all possible rules that could generate different recovery rates for the patients. Therefore, using single rule information, we cannot filter useful rules from all possible rules. And the study case aim is to investigate recurrent ovarian endometriomas under different conditions in a possible rule. For these reasons, we use *t*-test to verify rules to discover some useful description rules after all possible rules from the tree are generated. Then, we use *t*-test to verify all possible rules obtained from the decision tree to discover useful knowledge. Moreover, since the values of BMI and CA_125 may be missing, the number of patients of BMI or CA_125 attribute is not equal to 212.

A. Effect of CA_125 Value

The rule "Cyst_size < 4.25 and CA_125 ≥ 146.81 ==> Recovery = 0" should be ignored in the discussion because of the over-fit problem. However, there are 13 patients whose CA_125 value is greater than 146.81 in the whole training data. Therefore, we should check whether this CA_125 value presents some interesting descriptions about the recovery rate.

In this section, all patients are divided into two groups according to three different conditions, as shown in Table III. As shown by test 1 in Table III, all patients are divided into two groups according to the cut point of the CA_125 attribute. Group 1 contains patients whose CA_125 values are less than 146.81, while CA_125 values of patients in group 2 are not less than 146.81. From test 1 in the table, we conclude that when all patients are divided into two groups, the recovery rate of group 1 (89/189, 47.09%) and group 2 (1/13, 7.69%) is significantly different (p = 0.002769). From tests 2 and 3, we conclude that when a patient's CA_125≥146.81, there is no significant difference between the recovery rates of the two groups. In other words, when CA_125 ≥ 146.81, the recovery rate (or recurrence rate) will mainly be affected by this CA_125 value regardless of the cyst size and BMI value. That is, the recovery rate of that patient would be very low regardless of treatment type. These new useful descriptions cannot be discovered by the tree-generation algorithm directly. Actually, when a patient's CA_125 ≥ 146.81, the patient must have other diseases; otherwise, the value would not be greater than or equal to 146.81.

B. Effect of Cyst_size and BMI Values

As indicated in the above section, when CA_125 ≥ 146.81, the recovery rate (or recurrence rate) will mainly be affected by this CA_125 value regardless of the Cyst_size and BMI values. Therefore, we do not consider this attribute in the following discussion. According the cut points of the Cyst_size and BMI attributes, as shown in Table Ⅳ, all patients can be divided into two groups according to four different conditions. From Table Ⅳ, we conclude that the recovery rate (or recurrence rate) will mainly be affected by the cyst size.

TABLE IV.
RECOVERY RATES OF GROUPS

test	Group 1	Group 2	Group 1 Recovery-rate (y/total, rate)	Group 2 Recovery-rate (y/total, rate)	*p*-value
1	Cyst_size <4.25	Cyst_size ≥4.25	43/59. 72.88%	49/153, 32%	1.47E-08
2	BMI< 22.7516	BMI≥22.7516	75/155, 48.39%	9/26, 34.62%	0.097306
3	Cyst_size <4.25 and BMI< 22.7516	Cyst_size <4.25 and BMI≥22.7516	37/47, 78.72%	1/5, 20%	0.002115
4	Cyst_size ≥4.25 and BMI< 22.7516	Cyst_size ≥4.25 and BMI≥22.7516	38/108. 35.19%	8/21, 38.1%	0.400407

TABLE III.
T-TEST RESULTS FOR CA_125

test	Group 1	Group 2	Group 1 Recovery-rate (y/total, rate)	Group 2 Recovery-rate (y/total, rate)	*p*-value
1	CA_125<146.81	CA_125≥146.81	89/189, 47.09%	1/13, 7.69%	0.002769
2	CA_125≥146.81 and Cyst_size<4.25	CA_125≥146.81 and Cyst_size≥4.25	0/3, 0%	1/10, 10%	0.302958
3	CA_125≥146.81 and BMI< 22.7516	CA_125≥146.81 and BMI≥22.7516	1/8, 12.50%	0/4, 0%	0.252925

From test 2 in Table II, we conclude that, for patients whose cyst size is less than 4.25 and CA_125 is not less than 146.81, the BMI value, 22.7516, will significantly affect the recovery rate. However, we already know that when a patient's CA_125 is greater than 146.81, the recovery rate of the patient is very low. Therefore, the question in mind is that how about for those patients whose cyst size is less than 4.25. Will the value of BMI significantly affect the recovery rate of patients regardless of the CA_125 value? From test 3 in Table Ⅳ, we conclude that the recovery rates of group 1 (43/59, 72.88%) and group 2 (1/5, 20%) are significantly different (p = 0.002115). In other words, we can say that when the cyst size is less than 4.25, the BMI value will affect the recovery rate. We get a new useful description that cannot be discovered by the tree-generation algorithm directly. This new description can be interpreted as follows: "For those patients whose cyst size is less than 4.25, when the patient's BMI value is less than 22.7516, the recovery rate is 78.72%; otherwise, the recovery rate is 20% only. That is, for those patients whose cyst size is less than 4.25, when the BMI value is not less than 22.7516, the value of BMI will significantly affect the recovery rate."

C. Interpreting Mining Results with Ethanol Instillation

As pointed out by Noma and Yoshida [9], ethanol instillation into the cyst cavity for more than 10 min was most effective at reducing the recurrence rate. Therefore, another goal is to verify whether aspiration followed by the injection of 95% ethanol left in situ is an effective treatment for ovarian endometriomas under different conditions. Therefore, in the data preprocessing step, we classify the values of ethanol irrigation duration into two types: less than 10 min and retention. Therefore, for each condition, all patients are divided into two groups (group 1: ethanol irrigation; group 2: ethanol retention) by the resultant cut points. The values of *t*-test for the above conditions are shown in Table V. From the table, we give the following conclusions:

(1) When CA_125 > 146.81, recurrence rate is very high regardless of treatment type.

(2) When Cyst_size ≥ 4.25, the recovery rate of group 1 (11/49, 22.45%) and group 2 (38/104, 36.54%) is significantly different (p = 0.041180).

(3) Except the above two conditions, although the recovery rate of group 2 is better than that of group 1, there is no significant difference between these two groups.

TABLE V.
RECOVERY RATES OF ETHANOL IRRIGATION AND ETHANOL RETENTION GROUPS ABOUT DIFFERENT CONDITIONS

Condition	Group 1 Recovery-rate (y/total, rate)	Group 2 Recovery-rate (y/total, rate)	p-value
All patients	27/83, 36.99%	65/139, 46.76%	0.086980
CA_125 <146.81	24/59, 40.68%	65/130, 50%	0.118197
CA_125 ≥146.81	1/5, 16.67%	0/7, 0%	0.149845
Cyst_size <4.25	16/24, 66.67%	27/35, 77.14%	0.191314
Cyst_size ≥4.25	11/49, 22.45%	38/104, 36.54%	0.041180
BMI < 22.7516	21/46, 45.65%	54/109, 49.54%	0.330266
BMI≥22.7516	3/12, 25%	6/14, 42.86%	0.180010
Cyst_size <4.25 And BMI< 22.7516	12/17, 70.59%	25/30, 83.33%	0.157719
Cyst_size <4.25 And BMI≥22.7516	1/3, 33.33%	0/2, 0%	0.247512
Cyst_size ≥4.25 And BMI< 22.7516	9/29, 31.03%	29/79, 36.71%	0.294158
Cyst_size ≥4.25 And BMI≥22.7516	2/9, 22.22%	6/12, 50%	0.106864

Ⅴ. DISCUSSION AND CONCLUSION

Endometriosis is a complex disease with several attributes. Each of the attributes, such as the stage of the disease, symptoms of the disease, serum level of CA-125, can affect the making of treatment decisions and the results of the treatment.

This study integrates the statistical method and decision tree techniques to analyze the postoperative status of ovarian endometriosis patients. Experimental results show that some new interesting knowledge can be found by our approach. However, our method works only for a simple tree. To deal with more complex decision trees, we want to integrate other data mining the approaches to analyze postoperative status of ovarian endometriosis patients in the future. We hope to find more precise description knowledge about the recovery rates of the patients under different conditions.

REFERENCES

[1] . Tang, T.I., et al., "A Comparative Study of Medical Data Classification Methods Based on Decision Tree and System Reconstruction Analysis," Industrial Engineering and Management Systems, vol. 4, p. 102¡V108, 2005.

[2] . Cui, Z. and G. Zhang, "A novel medical image dynamic fuzzy classification model based on ridgelet transform," Journal of Software, vol. 5, pp. 458-465, 2010.

[3] . Scheetz, L.J., J. Zhang, and J. Kolassa, "Classification tree modeling to identify severe and moderate vehicular injuries in young and middle-aged adults," Artificial Intelligence in Medicine, vol. 45, pp. 1-10, 2009.

[4] . Ozkan, S. and A. Arici, "Advances in treatment options of endometriosis," Gynecol Obstet Invest, vol. 67, pp. 81-91, 2009.

[5] . Aboulghar, M.A., et al., "Ultrasonic transvaginal aspiration of endometriotic cysts: an optional line of treatment in selected cases of endometriosis," Human Reproduction, vol. 6, p. 1408, 1991.

[6] . Aboulghar, M.A., et al., "Treatment of recurrent chocolate cysts by transvaginal aspiration and

tetracycline sclerotherapy," Journal of assisted reproduction and genetics, vol. 10, pp. 531-533, 1993.
[7] . Mesogitis, S., et al., "Combined ultrasonographically guided drainage and methotrexate administration for treatment of endometriotic cysts," The Lancet, vol. 355, pp. 1160-1160, 2000.
[8] . Acien, P., et al., "GnRH analogues, transvaginal ultrasound-guided drainage and intracystic injection of recombinant interleukin-2 in the treatment of endometriosis," Gynecologic and obstetric investigation, vol. 55, pp. 96-104, 2000.
[9] . Noma, J. and N. Yoshida, "Efficacy of ethanol sclerotherapy for ovarian endometriomas," International Journal of Gynecology & Obstetrics, vol. 72, pp. 35-39, 2001.
[10] . Akamatsu, N., et al., "Ultrasonically Guided Puncture of Endometrial Cyst: Aspiration of Contents and Infusion of Ethanol," Acta Obstetrica et Gynaecologica Japonica, vol. 40, pp. 1214-1215, 1988.
[11] . Hsieh, C.L., et al., "Effectiveness of ultrasound-guided aspiration and sclerotherapy with 95% ethanol for treatment of recurrent ovarian endometriomas," Fertility and sterility, vol. 91, pp. 2709-2713, 2009.
[12] . Akan, E., et al., "Predictive Power of Activin A Levels in the Prognosis of First Trimester In Vitro Fertilization Pregnancies," Journal of Women's Health, 2011.
[13] . Acien, P., et al., "Use of intraperitoneal interferon [alpha]-2b therapy after conservative surgery for endometriosis and postoperative medical treatment with depot gonadotropin-releasing hormone analog: a randomized clinical trial* 1," Fertility and sterility, vol. 78, pp. 705-711, 2002.
[14] . Takano, M., et al., "The impact of complete surgical staging upon survival in early-stage ovarian clear cell carcinoma a multi-institutional retrospective study," International Journal of Gynecological Cancer, vol. 19, pp. 1353-1357, 2009.
[15] . Van Calster, B., et al., "Towards a clinical decision support system for pregnancies of unknown location," in *Proceedings - IEEE Symposium on Computer-Based Medical Systems*, Jyvaskyla, 2008, pp. 581-583.
[16] . Ben VAN CALSTE, "Predictive diagnostic models for gynecologic applications with focus on multi-class classification," PhD thesis, Dept. of Electrical Engineering, Katholieke Universiteit Leuven, Belgium, 2008.
[17] . Kyama, C.M., et al., "Evaluation of endometrial biomarkers for semi-invasive diagnosis of endometriosis," Fertility and sterility, vol. 95, pp. 1338-1343.e3, 2011.
[18] . Seeber, B., et al., "Panel of markers can accurately predict endometriosis in a subset of patients," Fertility and sterility, vol. 89, pp. 1073-1081, 2008.
[19] . Liu, H., et al., "Detection of endometriosis with the use of plasma protein profiling by surface-enhanced laser desorption/ionization time-of-flight mass spectrometry," Fertility and sterility, vol. 87, pp. 988-990, 2007.
[20] . Van Holsbeke, C., et al., "Endometriomas: Their ultrasound characteristics," Ultrasound in Obstetrics and Gynecology, vol. 35, pp. 730-740, 2010.
[21] . Bellazzi, R. and B. Zupan, "Predictive data mining in clinical medicine: Current issues and guidelines," international journal of medical informatics, vol. 77, pp. 81-97, 2008.
[22] . Kim, I.C. and Y.G. Jung, "Using Bayesian networks to analyze medical data," in *Lecture Notes in Artificial Intelligence (Subseries of Lecture Notes in Computer Science)*, Leipzig, 2003, pp. 317-327.
[23] . Hsieh, C.L., C.S. Shiau, and L.M. Lo, "Effectiveness of ultrasound-guided aspiration and sclerotherapy with 95% ethanol for treatment of recurrent ovarian endometriomas," Fertility and sterility, vol. 91, pp. 2709-2713, 2009.
[24] . Kinkel, K., et al., "Indeterminate ovarian mass at US: Incremental value of second imaging test for characterization-meta-analysis and Bayesian analysis," Radiology, vol. 236, pp. 85-94, 2005.
[25] . Wang, H. and P. Zhang, "A quantitative method for pulse strength classification based on decision tree," Journal of Software, vol. 4, pp. 323-330, 2009.
[26] . Breiman, L., et al., "Classi cation and regression trees," Wadsworth, Belmont, 1984.
[27] . Wang, K., S. Zhou, and Y. He, "Growing decision trees on support-less association rules," in *Proceeding of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Boston, MA, 2000, pp. 265-269.

Ming-Yang Chang, M.D. is currently working as an Associated professor in Gynecology, Chang Gung University School of Medicine, Taoyuan, Taiwan. His research interests include Infertility, endometriosis and female reproductive medicine.

Chien-Chou Shih received the Ph.D. degree in information engineering from Tamkang University, Taiwan, in 1998. He is currently an Assistant Professor with the Department of Information and Communication, Tamkang University. His research interests include embedded software programming, data mining applications, and engineering design education.

Dina-An Chiang received the BS degree in hydraulic engineering from Chung Yuan Christian University, Taiwan, in 1981, and the MS and PhD degrees in computer science from the University of Southwestern Louisiana in 1986 and 1990, respectively. He is currently a professor in the Department of Computer Science. His research interests include fuzzy, relational databases and data mining.

Chun-Chi Chen received the MS degrees in computer science and information engineering from Tamkang University in Taipei, Taiwan, in 2003. His research interests include relational databases and data mining.