Hindawi Publishing Corporation
Evidence-Based Complementary and Alternative Medicine
Volume 2013, Article ID 493626, 6 pages
http://dx.doi.org/10.1155/2013/493626

Research Article

Zheng Classification with Missing Feature Values Using Local-Validity Approach

Yan Wang[1,2] and Lizhuang Ma[3,4]

[1] *School of Continuing Education, Shanghai Jiao Tong University, Shanghai 200240, China*
[2] *Provincial Key Laboratory for Computer Information Processing Technology, Soochow University, Suzhou 215006, China*
[3] *Department of Computer Science & Engineering, Shanghai Jiao Tong University, Shanghai 200240, China*
[4] *Center of Traditional Chinese Medicine Information Science and Technology, Shanghai University of TCM, Shanghai 201203, China*

Correspondence should be addressed to Yan Wang; wangyan8383@sjtu.edu.cn and Lizhuang Ma; ma-lz@cs.sjtu.edu.cn

Received 26 July 2013; Revised 15 October 2013; Accepted 15 October 2013

Academic Editor: Shi-bing Su

Zheng classification is a very important step in the diagnosis of traditional Chinese medicine (TCM). In clinical practice of TCM, feature values are often missing and incomplete cases. The performance of Zheng classification is strictly related to rates of missing feature values. Based on the pattern of the missing feature values, a new approach named local-validity is proposed to classify zheng classification with missing feature values. Firstly, the maximum submatrix for the given dataset is constructed and local-validity method finds subsets of cases for which all of the feature values are available. To reduce the computational scale and improve the classification accuracy, the method clusters subsets with similar patterns to form local-validity subsets. Finally, the proposed method trains a classifier for each local-validity subset and combines the outputs of individual classifiers to diagnose zheng classification. The proposed method is applied to the real liver cirrhosis dataset and three public datasets. Experimental results show that classification performance of local-validity method is superior to the widely used methods under missing feature values.

1. Introduction

1.1. The Concept of Zheng Classification. Traditional Chinese medicine (TCM) is one of the most important complementary medicines used increasingly in the world [1]. Zheng classification enables the doctor to determine the stage that the disease developed and the location of the disease [2]. Zheng classification is the method of recognizing and diagnosing diseases by analyzing patient information based on TCM theories and the doctor's experiences [3].

In an attempt to achieve effective and objective standard of Zheng classification, various data mining approaches are used to construct the classifier on TCM dataset. Figure 1 shows the process of intelligent Zheng classification.

1.2. Missing Feature Values: The Literature Review. In clinical practice of traditional Chinese medicine, feature values are often missing and incomplete cases. Missing feature values could be caused by various reasons, such as error of data measure, error of data understanding, erroneous human imputation, or restriction of data collecting [4, 5]. The performance of intelligent Zheng classification model in TCM is strictly related to the rate of missing feature values, but most common methods are short of the ability to solve the missing feature problem [4–6].

At present, the most common strategy for dealing with absent values is essentially to ignore them [7]. The cases with missing feature values are deleted before constructing the Zheng classification model [7]. Although improving the classification performance in some degree, deletion may discard some important information within the missing feature values, especially under the condition of insufficient TCM data. So deleting the data with missing feature values directly is difficult to meet the TCM clinical application.

Considering the shortcomings of the deletion method, imputation solution comes into being. Imputation is the substitution of a missing feature value with a meaningful estimate. Evidence theory is used to predict the missing feature values [8]. However, the evidence function should be learned in advance. The literature [9, 10] fills the missing feature by

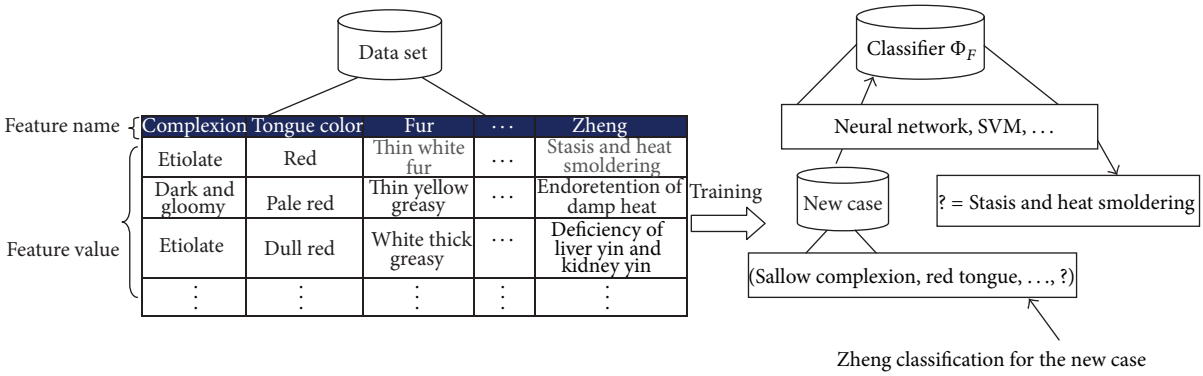

Figure 1: The process of intelligent Zheng classification.

statistics method and Bayesian model, respectively. Nevertheless, these methods need to know probability distribution, which is difficult to be acquired in fact. In some applications, expert experience could be used to form the complete feature values. However, the prediction method for missing data by experts is subjective.

In recent decades, data mining imputation methods are beginning to attract much attention [11]. Logistic regression [12], subspace [4], neural network [13], and rough sets theory [14] have been applied to deal with missing feature values. These methods construct a predictive model to estimate the missing feature values from information within cases. However, imputation method will introduce new noise into cases, and the classification accuracy will decrease subsequently.

When dealing with the missing feature values, deletion and imputation methods will change the original dataset more or less. To avoid the problem of deletion and imputation methods, the literature [15–17] presents a selective Bayes classifier for classifying missing values with a simpler formula for computing gain ratio. Nevertheless, the method needs to satisfy the premise that features should be independent of each other. In TCM clinical practice, it is difficult to guarantee the characteristics of independence.

To overcome the limitation of the methods mentioned above, the proposed local-validity approach need not estimate the missing feature values or remove the deficient cases. It focuses on constructing intelligent Zheng classifier on the original cases directly. Firstly, the method finds the local-validity subset (LVS) within dataset and constructs the Zheng classifier on each LVS. Finally, the performance of each individual classifier is assessed and combined depending on the classification matrix to estimate the final output.

The rest of the paper is organized as follows. Section 2 describes the dataset and the ideas of the proposed local-validity method. The experimental results based on the method are shown in Section 3. Finally, conclusion is given in Section 4.

2. Material and Methods

2.1. Description of Dataset. 153 liver cirrhosis cases with three different Zheng classifications (i.e., stasis-heat smoldering zheng, damp-heat smoldering zheng, and liver-kidney yin deficiency zheng) have been collected from Shanghai University of Traditional Chinese Medicine. The dataset includes 52 cases with stasis-heat smoldering zheng, 61 cases with damp-heat smoldering zheng and 40 cases with liver-kidney yin deficiency zheng. Each case includes 40 TCM features selected by clinicians as the significant factors to identify the liver cirrhosis zheng.

Features are encoded using the four-value ordinal scales measured by the severity degree:

(i) 1 representing no corresponding symptoms;

(ii) 2 for the normal level;

(iii) 3 for the medium serious level;

(iv) 4 representing the most serious.

Among all features, twenty-three features are missing in varying degrees. In this paper, the missing percentage α is defined as

$$\alpha = \frac{|U'|}{|U|}, \tag{1}$$

where $|U'|$ denotes the number of cases with missing feature values and $|U|$ denotes the total number of cases.

The list of these features and the corresponding missing percentage are shown in Table 1.

2.2. The Proposed Local-Validity Approach. As mentioned above, the local-validity idea overcomes the limitations discussed in the previous section. The flowchart of the proposed approach is shown in Figure 2. The subsequent subsections are organized as follows. First, the TCM zheng classification system with missing feature values is defined. Then, we describe how LVS is selected and how the individual classifier is trained on every LVS. Finally, we present how the individual classification results are combined to boost up the classification performance.

2.2.1. Definition of Zheng Classification System. Zheng classification system with missing feature values in TCM can be viewed as a 3-tuple $S = \langle U, F, g \rangle$, where U is a nonempty finite set of cases and F is a nonempty finite set of features.

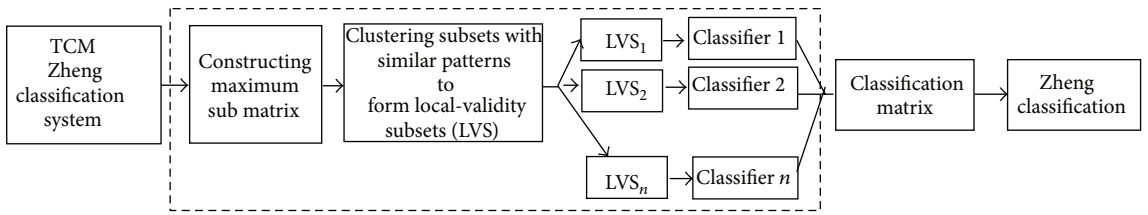

Figure 2: The overall view of the proposed local-validity approach.

Table 1: Description of liver cirrhosis TCM dataset used in the experiment.

Feature name	α (%)
(1) Lassitude and fatigue	5
(2) Head heaviness	0.1
(3) Spontaneous sweat	10
(4) Nocturnal polyuria	0
(5) Depression	4
(6) Gingiva bleeding	0.2
(7) Blurred vision	3.1
(8) Reduced appetite	0
(9) Dry and bitter taste	0
(10) Abdominal pain	1
(11) Rib-side and flank distention and pain	2.3
(12) Low limbs puffy swelling	1.2
(13) Belching	0
(14) Yellow urine	0
(15) Scant urine	3.2
(16) Night sweat	1.1
(17) Sloppy stool	0.2
(18) Skin itching	2.1
(19) Skin bleeding	0
(20) Insomnia	0.3
(21) Limp aching lumbar and knees	0
(22) Tinnitus	0
(23) Hypochondriac distending pain	3.8
(24) Abdominal distension	0.1
(25) Yellow body	0
(26) Acid regurgitation	0
(27) Liver palm	0
(28) Dazzle	0.1
(29) Chill and cold limbs	2.1
(30) Constipation	0
(31) Vexing heat in the five heart	0.3
(32) Nose bleeding	0
(33) Rashness impatience and irascibility	0
(34) Fatigued and heavy limbs	0
(35) Dry eyes	4.1
(36) Epigastralgia	0.1
(37) Foul breath	0.2
(38) Yellow eyes	0
(39) Nausea vomit	0
(40) Spider naïve	0.1

Table 2: A dataset with missing feature values.

U	f_1	f_2	f_3	f_4
x_1	$*$	$*$	1	0
x_2	1	1	$*$	$*$
x_3	0	1	1	0
x_4	1	0	$*$	1

For $\forall f \in F$ and $x \in U$, $g(x, f)$ denote the value that x holds on feature f. Then, in zheng classification system with missing feature values, $\exists x \in U$ and $\exists f \in F$ that satisfies $g(x, f) = *$. Here, we assume that the missing feature values are denoted by "$*$."

An example of zheng classification system with missing feature values is shown in Table 2.

2.2.2. Finding Local-Validity Subsets. It is common that the number of missing feature values is n ($n \geq 1$) in TCM clinical application. Based on the maximum sub-matrix theory, the missing feature values are considered as barrier points. The local-validity approach enumerates the maximum feature vector with complete values. Thus, the proposed method starts with a binary matrix M whose element is defined as

$$M_{i,j} = \begin{cases} 1, & g\left(f_i, x_j\right) \neq *, \\ 0, & g\left(f_i, x_j\right) = *. \end{cases} \tag{2}$$

The element $M_{i,j} = 0$ if the ith feature is missing in the jth case.

The matrix M of the dataset presented in Table 2 is given as follows:

$$M = \begin{bmatrix} 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 \end{bmatrix}. \tag{3}$$

Matrix M finds the maximum feature vector (MFV) that covers the most complete data. Each MFV identifies a local-validity pattern P; the formula of P is as follows:

$$\exists x \in U, \quad \forall f_i \in \overline{P} \wedge \forall f_j \in \left(F - \overline{P}\right), \tag{4}$$

satisfing $(g(f_i, x) = *) \wedge (g(f_j, x) \neq *)$ $(i \neq j)$.

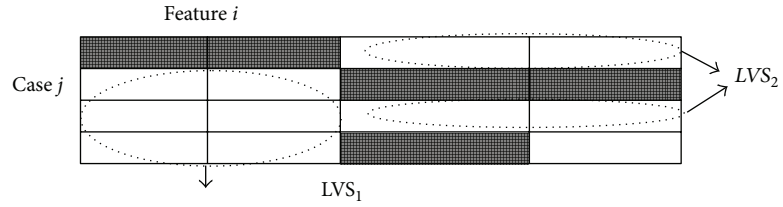

Figure 3: Example of local-validity subset.

Thus, the corresponding local-validity pattern P corresponding to Table 2 is

$$P_1 = \{f1, f2\}, \quad P_2 = \{f3, f4\}, \quad P_3 = \{f1, f2, f3, f4\}, \quad P_4 = \{f1, f2, f4\}. \tag{5}$$

Each pattern P maps the corresponding LVS.

LVS is a collection of the cases that have no missing values for a specific feature subset and the collection of LVS includes all of the cases in the original data. The formula of LVS is as

$$(\forall x \in \mathrm{LVS}_k) \wedge (\forall f_i \in \mathrm{LVS}_k), \tag{6}$$

satisfing $g(f_i, x) \neq *$, where LVS_k represents the kth local-validity subset.

Four LVS can be found from the dataset presented in Table 2 and Figure 3 shows two of them.

The process of finding local-validity subset can be described as follows.

(1) For an original given dataset, generate matrix M.

(2) In feature space, traverse matrix M to generate maximum feature vector.

(3) For each feature vector, find the corresponding LVS.

As the feature missing percentage ascends, the number of LVS will increase. The large number of LVS will affect the computation complexity. Then, this problem will be translated into a clustering problem. LVS with the similar pattern will be merged.

2.2.3. Clustering Local-Validity Subsets with Similar Patterns. It is desirable to cover the entire data with as few local-validity subsets as possible and obtain the overall best performance.

The preliminary research results [17] show that there are inherent consistencies between mutual information and subset aggregate.

Considering the cross entropy between two local-validity subsets,

$$\mu_{i,k} = -\log\left(\frac{1}{N_i N_k}\sum_{p=1}^{N_i}\sum_{q=1}^{N_k} G\left(x_p - y_q, \sigma^2\right)\right), \tag{7}$$

where G is the Gaussian kernel and σ^2 is the variance of Gauss function. N_i is the number of cases in subset LVS_i and N_k represents LVS_k.

Table 3: Performance comparison of three methods on liver cirrhosis dataset.

Classification accuracy (%)		
Deletion	Imputation	Local-validity
68.67	70.67	**80.33**

The bold values are used to emphasize the best Zheng classification performance.

Then, the mutual information $I(i, k)$ between LVS_i and LVS_k can be defined as follows:

$$I(i,k) = \mu_{i,k} \log \frac{N_{i,k}}{N_i + N_k}, \tag{8}$$

where $N_{i,k}$ represents the number of cases that belongs to two subsets at the same time.

The larger $I(i, k)$ is, the stronger the correlation degree between LVS_i and LVS_k is. Based on k-nearest neighbor algorithm, the subset with strong correlation degree is clustered to form a new subset. In this paper, the kth cluster is represented by a set of LVS indices Ω_k.

2.2.4. Constructing the Zheng Classification Matrix. Once LVS is chosen, an individual zheng classifier is needed for each Ω_k. In TCM zheng classification previous studies [18], the zheng classification matrix is proposed to merge the outputs of multizheng classifiers under the complete dataset.

Under missing feature values, in order to boost up the zheng classification performance, the complete degree λ_i is introduced into the zheng classification matrix Y to estimate the final output. Then, zheng classification matrix Y is updated as

$$Y = \arg\max_{w \in Y} \sum_{i=1}^{k} \lambda_i G_i^w, \tag{9}$$

where G_i^w represents the performance that a new case is diagnosed as w under the Ω_i local-validity subset.

3. Experimental Results

3.1. Local-Validity versus Other Methods on Liver Cirrhosis Dataset. To evaluate the performance of the proposed method, we carried out experiments on a real TCM liver cirrhosis dataset with missing data. Description of the dataset is presented in Table 1.

To analyze the improvement in zheng classification accuracy, three different methods are used to deal with the missing values.

Table 4: The performance comparison of three methods on three public datasets.

(a) Lymphography

α	Diagnosis accuracy (%)		
	Deletion	Imputation	Local-validity
0	**85.14**	**85.14**	83.7
0.05	**85.82**	83.78	84.02
0.10	81.21	81.08	**92.16**
0.20	78.92	77.43	**88.11**

(b) SPECT heart

α	Diagnosis accuracy (%)		
	Deletion	Imputation	Local-validity
0	**82.4**	**82.40**	82.05
0.05	**82.28**	82.02	83.06
0.10	82	80.52	**85.25**
0.20	80.2	78.08	**86.06**

(c) Lung cancer

α	Diagnosis accuracy (%)		
	Deletion	Imputation	Local-validity
0	**90.6**	**90.6**	90.05
0.05	89.25	**90.12**	89.17
0.10	86.23	83.37	**87.29**
0.20	81.67	82.02	**82.37**

The bold values are used to emphasize the best Zheng classification performance.

The zheng classification accuracy is first estimated by simply removing the cases with missing values. Then, mean value imputation method is applied to impute missing feature values. Finally, the proposed local-validity approach is applied on the original dataset directly.

Considering the liver cirrhosis data is not sufficient, ten times 10-fold cross-validation is used for the assessment of classification performance. In cross validation, the data is split into ten approximately equal partitions and each in turn is used for testing and the remainder is used for training. That is, use nine-tenths for training and one-tenth for testing and repeat the procedure ten times so that, in the end, every case has been used exactly once for testing [19].

To get a reliable error estimate, the cross-validation process is repeated for 10 times, and the results are averaged [19].

The average classification accuracies are listed in Table 3. The best performance is emphasized using a boldfaced font.

As seen in Table 3, the performance of local-validity approach outperforms the deletion and imputation methods on liver cirrhosis dataset.

It should be pointed out that there are 23 feature values missing in original 40 features. Simply, deletion may introduce substantial biases, and imputation will introduce noise. With the increase of missing rate, problems of deletion and imputation will be more obvious. On the other hand, local-validity method constructs the zheng classification on the original dataset directly. The method can avoid the noise and biases problems.

3.2. Local-Validity versus Other Methods on Other Datasets. We also do experiments on three public datasets: lymphography, SPECT heart, and breast cancer.

Because these three datasets are complete, the diagnosis performance can be evaluated effectively. We replace randomly the feature value with "$*$" based on different missing percentages $\alpha = \{0, 0.05, 0.10, 0.20\}$ in these three public datasets. The results are shown in Table 4.

From Tables 4(a), 4(b), and 4(c), it can be seen that the performance of local-validity method is lower than that of deletion and imputation with $\alpha = 0.05$. With $\alpha = 0.1$ and $\alpha = 0.2$, the proposed method performs well than other methods on three datasets. This shows that the performance of local-validity method is more stable than that of the other two methods, and the effect will be more obvious with the number increment of the missing cases.

In summary, the proposed local-validity algorithm is applicable to the dataset with small number of cases and a large percentage of missing values.

4. Conclusions

Although various machine-learning algorithms have been used to construct the zheng classification model in TCM, most of them deal with complete feature values. In fact, missing feature values are inevitable in TCM clinical application. Therefore, methods of constructing zheng classifier for missing data deserve more attention.

By analyzing missing data processing methods, this paper presents a local-validity approach for zheng classification with missing feature values. The proposed approach contains the following characteristics.

(1) Instead of deleting or imputing the absent values, the proposed approach discovers the local-validity subsets from the original cases. Therefore, the proposed approach avoids the introduction of noise data.

(2) Our method constructs zheng classifier on the original dataset directly and needs no assumption about the missing mechanism.

(3) During the local-validity subset discovery phase, the formula for computation of the local-validity subset is presented. Then, the zheng classification matrix is described to combine the classification results of multi-individual classifiers.

(4) Through experiments, we can conclude that the proposed method is an appropriate solution to missing feature values problems in TCM zheng classification. The results show that the proposed approach outperforms the deletion and imputation methods as the amount of missing feature values increases.

(5) Further research is under way concerning the relationship between the scale of local-validity subset and classification accuracy in order to get the optimum diagnostic result.

Conflict of Interests

The authors declare that they have no conflict of interests.

Acknowledgments

The research is sponsored by Open Research Fund of Jiangsu Provincial Key Laboratory for Computer Information Processing Technology (no. KJS1226) and Biomedical Engineering Cross Funds of Shanghai Jiao Tong University (no. YG2012MS28). The authors gratefully acknowledge all the researchers from the Shanghai University of Traditional Chinese Medicine for the TCM databases.

References

[1] S. K. Pal, "Complementary and alternative medicine: an overview," *Current Science*, vol. 82, no. 5, pp. 518–524, 2002.

[2] B. Flaws and P. Sionneau, *The Treatment of Modern Western Medical Diseases with Chinese Medicine*, Blue Poppy Press, 2005.

[3] J. Si, L. Sun, N. Dai et al., "Study of sEGF level in chronicatrophic gastritis with either Chinese traditional medicine or Western medicine," *Journal of Zhejiang University Science*, vol. 3, no. 2, pp. 243–246, 2002.

[4] L. Nanni, A. Lumini, and S. Brahnam, "A classifier ensemble approach for the missing feature problem," *Artificial Intelligence in Medicine*, vol. 55, no. 1, pp. 37–50, 2012.

[5] R. Polikar, J. DePasquale, H. S. Mohammed, G. Brown, and L. I. Kuncheva, "Learn++.MF: a random subspace approach for the missing feature problem," *Pattern Recognition*, vol. 43, no. 11, pp. 3817–3832, 2010.

[6] G. Z. Li, S. X. Yan, M. Y. You et al., "Intelligent ZHENG classification of hypertension depending on ML-kNN and information fusion," *Evidence-Based Complementary and Alternative Medicine*, vol. 2012, Article ID 837245, 5 pages, 2012.

[7] H.-B. Qu, L.-F. Mao, and J. Wang, "Method for self-extracting diagnostic rules of blood stasis syndrome based on decision tree," *Chinese Journal of Biomedical Engineering*, vol. 24, no. 6, pp. 709–727, 2005.

[8] M. A. Boujelben, Y. D. Smet, A. Frikha, and H. Chabchoub, "Building a binary outranking relation in uncertain, imprecise and multi-experts contexts: the application of evidence theory," *International Journal of Approximate Reasoning*, vol. 50, no. 8, pp. 1259–1278, 2009.

[9] A. S. Salama, "Topological solution of missing attribute values problem in incomplete information tables," *Information Sciences*, vol. 180, no. 5, pp. 631–639, 2010.

[10] Z. Huang, J. Li, H. Su, G. S. Watts, and H. Chen, "Large-scale regulatory network analysis from microarray data: modified Bayesian network learning and association rule mining," *Decision Support Systems*, vol. 43, no. 4, pp. 1207–1225, 2007.

[11] M. Ghannad-Rezaie, H. Soltanian-Zadeh, H. Ying, and M. Dong, "Selection-fusion approach for classification of datasets with missing values," *Pattern Recognition*, vol. 43, no. 6, pp. 2340–2350, 2010.

[12] D. Williams, X. Liao, Y. Xue, L. Carin, and B. Krishnapuram, "On classification with incomplete data," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 3, pp. 427–436, 2007.

[13] I. A. Gheyas and L. S. Smith, "A neural network-based framework for the reconstruction of incomplete data sets," *Neurocomputing*, vol. 73, no. 16–18, pp. 3039–3065, 2010.

[14] X.-Y. Shao, X.-Z. Chu, H.-B. Qiu, L. Gao, and J. Yan, "An expert system using rough sets theory for aided conceptual design of ship's engine room automation," *Expert Systems with Applications*, vol. 36, no. 2, pp. 3223–3233, 2009.

[15] J. Chen, H. Huang, F. Tian, and S. Tian, "A selective Bayes Classifier for classifying incomplete data based on gain ratio," *Knowledge-Based Systems*, vol. 21, no. 7, pp. 530–534, 2008.

[16] F. Smeraldi, M. Defoin-Platel, and M. Saqi, "Handling missing features with boosting algorithms for protein-protein interaction prediction," *Lecture Notes in Computer Science*, vol. 6254, pp. 132–147, 2010.

[17] Y. Wang, Y. Gao, R. Shen, and F. Yang, "Selective ensemble approach for classification of datasets with incomplete values," *Advances in Intelligent and Soft Computing*, vol. 122, pp. 281–286, 2011.

[18] Y. Wang, L. Ma, and P. Liu, "Feature selection and syndrome prediction for liver cirrhosis in traditional Chinese medicine," *Computer Methods and Programs in Biomedicine*, vol. 95, no. 3, pp. 249–257, 2009.

[19] I. H. Witten and E. Frank, *Data Mining Practical Machine Learning Tools and Techniques*, Morgan Kaufmann Press, San Francisco, Calif, USA, 2005.