A Quantitative Method for Pulse Strength Classification Based on Decision Tree

Huiyan Wang
College of Computer Science and Information Engineering, Zhejiang Gongshang University, Hangzhou, China
cederic@mail.zjgsu.edu.cn

Peiyong Zhang*
Institute of VLSI Design, Zhejiang University, Hangzhou, China
E-mail: zhangpy@vlsi.zju.edu.cn

Abstract—**Pulse diagnosis is one of the most important examinations in Traditional Chinese Medicine (TCM). In response to the subjectivity and fuzziness of pulse diagnosis in TCM, quantitative systems or methods are needed to modernize pulse diagnosis. In pulse diagnosis, strength is one of the most difficult factors to recognize. To explore the quantitative recognition of pulse strength, a novel method based on decision tree (DT) is presented. The proposed method is testified by applying it to classify four hundreds pulse signal samples collected from clinic. The results are mostly accord with the expertise, which indicate that the method we proposed is feasible and effective and can identify pulse signals accurately, which can be expected to facilitate the modernization of pulse diagnosis.**

Index Terms—**pulse signal identification; decision tree; feature selection; quantitative diagnosis**

I. Introduction

Pulse diagnosis is one of the four examinations, namely inspection, ‘auscultation and olfaction’, inquiry and palpation. Doctors diagnose the patient by feeling the pulse beating at the measuring point of the radial artery, which requires long experiences and a high level of skill. Traditional pulse diagnosis is subjective and deficient in quantitative criteria of diagnosis, which affects the reliability and repeatability of pulse diagnosis. Therefore, quantitative methods are needed to classify pulse signal. A lot of effort is being spent on pulse signal analysis [1-7]. In Traditional Chinese Medicine (TCM), pulse signals are considered carrying important information that can reflect the health state of human body. The identification of pulse signals is the purpose of pulse diagnosis in TCM. Much work has been reported recently in this field, in which multivariable statistical analysis (MSA) was mostly utilized to construct pulse diagnostic models. On one hand, these methods need to determine the thresholds of pulse parameters. The thresholds are determined mostly through experiments, which are often unreliable and difficult to operate. On the other hand, MSA is a linear model, which cannot reflect the complex relationships between pulse signal and pulse type. In order to meet the requirement of nonlinearity in pulse diagnosis, many studies have been carried out to construct models for pulse recognition. Previous work, such as [3] and [4], built pulse signal classification system based on artificial neural network, in which the features used were very simple and not enough for pulse recognition of the complex pulse signals. In pulse diagnosis, time-domain parameters can reflect the specificity of pulse signals. So they are endowed with important physiological significance by specialists of Traditional Chinese Medicine (TCM) and have obvious medical diagnostic importance [7]. A study on the construction of pulse diagnostic model based on time-domain characteristic parameters was done by [5], which demonstrated that time-domain characteristics can be representative of pulse signals. In pulse diagnosis, strength is one of the most difficult factors to recognize. Pulse strength (PS) is the synthetical reflection of pulse force and its changing tread, and is hard to be represented by one or several characteristic parameters. Accordingly, the recognition of PS is more complicated. Up to now, Little research has been conducted in the identification of PS.

In our pioneer work [5], we constructed a pulse diagnostic model based on Bayesian networks (BNs), in which time-domain characteristic parameters were utilized and the predictive accuracy rate (PAR) of PS attained 89.74%, which was not satisfactory. On one hand, the errors may be resulted in by some factors, such as the dataset is imbalanced and the discretization method is not very suitable for PS. On the other hand, in BNs, the causal relationships present in graphics mode and the diagnostic rules cannot be induced directly. To get explicit rules for diagnosis and explore new way to recognize PS, a novel method based on decision tree (DT) is proposed in this study.

Project supported by the Scientific Research Fund of Zhejiang Provincial Education Department (Grant No. 20070593), the Major Program of Zhejiang Science and Technology Plan Item (Grant No. 2008c14100), Zhejiang Natural Science Foundation of China (Grant No. Y1080230).

* Corresponding author.

Decision tree (DT) is one of approaches to multistage decision making, and has been used for efficient acquisition of knowledge from mass amount of data. The basic idea of DT is to break up a complex decision into a union of several simpler decisions, hoping the final solution obtained this way would resemble the intended desired solution [8]. In this paper, DT is employed to classify PS on the basis of time-domain characteristic parameters. The time-domain parameters can reflect the specificity and be representative of pulse signals. So they are endowed with important physiological significance by specialists of TCM and have obvious medical diagnostic importance [6-7]. DT can be built off-line from a training dataset [11]. Some well-known DT generators, such as ID3 [9], GID3 [10], GID3* [11], and C4.5 [12], have been developed as methods of machine learning, among which C4.5 is known to be the most frequently used DT generator. Firstly, it is an improved model of ID3, and always used as a reference benchmark for the study and analysis of classification problems. Secondly, precious work [13] showed that it provides good classification ability and run fast. Thirdly, in the procedure of C4.5, the discretization and selection of attributes were performed jointly, which seems to be more suitable because the selection of an attribute is in fact a selection of one of differently discretized attributes. Thus, we used C4.5 combined with a normalized information gain [14] to build our model. First, the time-domain characteristic parameters were extracted. Second, the imbalanced dataset was corrected by using under-sampling the majority class technique. Third, the classification model of pulse signal was constructed based on DT. The performance of the model is validated by experiments. The results show that the scheme we proposed is feasible and can classify pulse signal accurately, which can be expected to be useful in the modernization of TCM.

II. Time-Domain Characteristic Parameters of Pulse Signal

Fig. 1 presents a period of a pulse waveform of a health volunteer, which is obtained by a pulse transducer. This pulse signal sample is a triple humped waveform, where S, P, E, K, F and G are the characteristic points. One period of pulse waveform is usually composed of three waves: percussion wave, tidal wave and dicrotic wave (Fig.1), which are three separate waves, base on which the parameters are extracted. The y-axis is the amplitude of the pulse signal, whose unit is gram force (g). The x-axis is the time, whose unit is millisecond. Fig. 2 illustrates the pulse signal acquisition system. The sampling rate is 100 Hz. The pulse transducer is belt-mounted and fixed on the radial pulse at the wrist when sampling pulse signal. The pressure can be regulated gradually from 0g to 250g through a vertical position regulator screw. As the contact pressure of pulse transducer increases, the amplitude of the pulse signal first increases, reaching a maximum point and then decreases. The output dynamic range of the pulse signal acquisition system is from zero to fifty gram force (g).

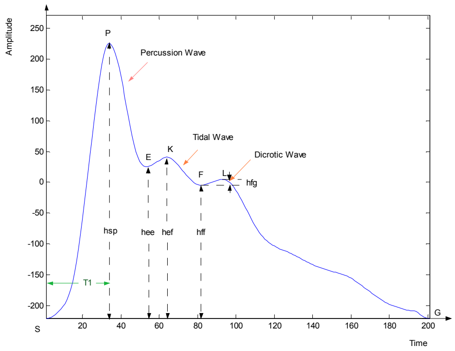


Figure 1. Time-domain parameters of pulse signal.

The main time-domain parameters are $h_{sp,}$ $h_{ee,}$ $h_{ef,}$ $h_{ff,}$ $h_{fg,}$ $T_{1},$ r_{fp}, r_{es} and r_{fs}. Thereinto, $r_{es} = h_{ef} / h_{sp}$, $r_{fp} = h_{ff} / h_{sp}$ *and* $r_{fs} = h_{fg} / h_{sp}$, which all have important physiological, pathologic and psychological significance, and have been testified to be important for diagnosis [6-7]. For example, the parameter r_{es} reflects the resilience and peripheral resistance of vascular wall. The parameter h_{ee}, which has same physiological signification with $h_{ef,}$ is not considered in this paper. To identify PS, we compute all the time-domain characteristic parameters and assume that these nine parameters are coequally important and all contribute to PS diagnosis. The detail procedure of feature extraction is reported in our pioneer work [1, 6].

III. The Category of PS

According to PS, pulse signal can be sorted into normal strength pulse (NS-pulse), replete pulse (R-pulse) and feeble pulse (F-pulse). Fig. 2(a) shows a R-pulse sample, the characteristic of which is that it can be felt vigorously and forcefully on both light and heavy pressure is named [7], while F-pulse is a pulse that is felt feeble and void, occurring when qi and blood are deficient or body fluid is impaired. Fig. 2(b) is a F-pulse sample. R-pulse is common in patients with deficiency syndrome, while F-pulse in ones with excess syndrome. The image features of F-pulse are that the length is short, the amplitude of percussion wave is small, the slope of ascending branch and descending branch is small, or the wave amplitude is moderate, but the dicrotic wave is relatively low or the curvature of descending branch is large [7]. The characteristics of R-pulse image are that the width and length are larger than normal, the percussion wave amplitude is large and wide, the ascending branch and descending branch slope is large and the wave canyon is relative high [7].

IV. Dealing With Class Imbalance

A dataset is imbalanced if the classes are not approximately equally represented [15], which is prevalent in many real-world applications, such as medical diagnosis. The methods to deal with class imbalance can be classified into two categories. One is to assign distinct costs to training examples [16]. The other is to resample the unbalanced dataset, either by over-sampling the minority class (OSMC) or under-sampling the majority class (USMC). It has been proved that USMC enables better classifiers to be built than OSMC, and a combination of the two cannot lead to classifiers that outperform those built utilizing only OSMC in most cases [17]. Thus USMC is selected in this work. The procedure is listed as follows [18]:

Step 1: Assume D_1 be the original dataset, D_2 be the set contains all minority class cases from D_1 and a randomly selected majority class cases subset.

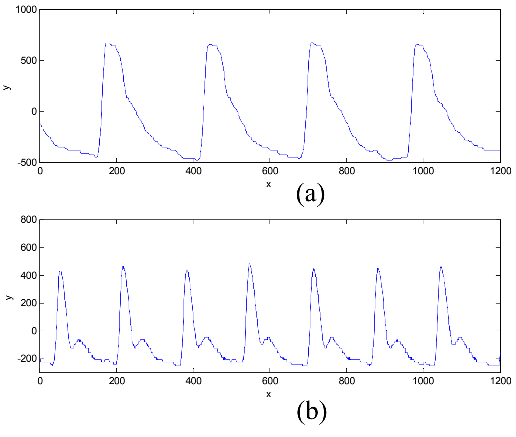
Step 2: Use the cases in D_2 to build a DT and employ it to classify D_1. Move all misclassified cases into D_2.

Step 3: Suppose c_1 and c_2 be two majority class cases in D_2 and each has a different class label. $H(c_1,c_2)$ denotes the distance between c_1 and c_2. If there is no case c_3 satisfies that $H(c_1,c_3) < H(c_1,c_2)$ or $H(c_2,c_3) < H(c_1,c_2)$, delete c_1 and c_2 from D_2. Then the resulting set is the required dataset.

V. Parameters Selection

The identification model is an classifier that presents the generalize relationship between the input attributes and target attribute. It is well known that redundant and irrelevant input attributes may degrade the performance of DT algorithm. Therefore, we build the diagnostic model combining with feature pre-selection method. As comparison, a novel mutual information based symptom selection algorithm (MISS) [22] and a symptom selection strategy based on BNs (MBSS) [23] are used in this study.

To reduce the irrelevant input attributes and improve the prediction accuracy of diagnostic model, feature selection is requisite. Many feature selection methods, such as filters [24] and wrappers [25], have developed. In order to achieve simplicity and scalability, many researchers use variable ranking as a baseline method for variable selection, among with mutual information based feature selection (MIFS) [26] is one of the most effective algorithms. MIFS algorithm selects an informative subset to be used as input data for the model to be built on the basis of the mutual information criterion.



(a)

(b)

Figure 2. Pulse sample
(a)R-pulse, (b) F-pulse

The result of symptoms selection using MIFS algorithm are influenced directly by a parameter β, which is usually determined through try and error method. In our previous work [22], we determine β based on the expertise (prior knowledge) of TCM and propose MISS algorithm, which is inspired by the idea of image template matching. MISS algorithm is validated to be superior to MIFS. In this study, we use MISS as one feature selection method and the gained variables are denoted as symptom set V_1.

In our another work [23], a Bayesian network structure learning algorithm is utilized to find out the most important symptoms that are interdependent with the disease, which is a supervised feature selection method and we called it MBSS. Firstly, all symptoms are evaluated and ranked by significance index S(F,C), which is defined as

$$S(F,C) = \frac{I(F,C)}{I_0} \tag{1}$$

where F is symptom and C is disease or syndrome, respectively. Suppose f and c is the events of F and C, then I(F,C) denotes the mutual information between variables c and f. I_0 is the prior entropy of C. If $S(F,C) < \sigma$, the symptom is deleted. Then the residual symptoms are selected as feature set V_0, which are exploited to learn BNs. Secondly, GBPS algorithm [27] is employed to learn a Bayesian network structure based on V_0. Thirdly, Markov blanket [28] is used to select symptoms. A Markov blanket is the minimum set of nodes that renders node X conditionally independent of all other nodes in the directed graph. The Markov blanket of a node X consists of the parents of X, the children of X, and the parents of the children of X. We can get rid of all nodes outside Markov blanket of X to obtain simplified BNs without influencing predictive accuracy rate. MBSS is employed as the second feature pre-selection algorithm in this work and the variables in Markov blanket are selected as symptom set V_2.

VI. Decision Tree Algorithm

A DT is a graphical representation of a procedure for classifying or evaluating an item of interest [19]. A tree has three types of nodes: a root node, which has no incoming edges and zero or more outgoing edges, internal nodes, each of which exactly one incoming edge and two or more outgoing edges, and leaf or terminal nodes, each of which has exactly one incoming edge and no outgoing edges. Root node and internal nodes contain attribute test conditions to separate cases. Each leaf node is assigned a class label. The C4.5 algorithm builds a DT using a divide and conquer (DC) strategy. Let a sample set D be composed of g classes c_1, c_2, …, c_g, having probabilities $p_1, p_2, \ldots, p_g$, respectively.

Let D_n be the training subset that is associated with note n. Then the procedure can be described as follows [13-14]:

Step 1: Compute the weighted frequency $freq(C_i, D_n)$ in D_n whose class label is C_i for $i \in [1, g]$.

Step 2: If all samples in D_n belong to a same class C_j or the number of cases in D_n is less than a threshold T_1, then the node n is a leaf and labeled as C_j.

Step 3: If D_n contains samples belong to two or more classes, then the information gain of each attribute is calculated. Let an attribute x divide D into k disjoint subsets $D_1, D_2, \ldots, D_k$.

The entropy of D is defined as

$$E(D) = -\sum_{i=1}^{g} p_i \log_2 p_i \tag{2}$$

Then the entropy $E(x, D)$ of D partitioned by x is denoted as

$$E(x,D) = -\sum_{i=1}^{k} P_i \sum_{j=1}^{g} p_{ij} \log_2 p_{ij} \tag{3}$$

where p_{ij} is the size of samples of class c_j in D_i relative to the size of D_i and P_i is the size of D_i relative to the size of D. Then the information gain is computed by

$$\text{gain}(x,D)=E(D)-E(x,D) \tag{4}$$

In general, an attribute may cause n-ary partition of the samples. Thus, in order to better accommodate the characteristics of the attribute in classification, a normalized gain is defined as [14]

$$NE(x,D) = \sum_{j=1}^{g} p_j \log_k p_j - \sum_{i=1}^{k} P_i \sum_{j=1}^{g} p_{ij} \log_k p_{ij} \tag{5}$$

which can also be expressed as

$$NE(x,D) = \frac{gain(x,D)}{\log_2 k}, k \geq 2 \tag{6}$$

The attributes with the highest $NE(x,D)$ is selected for the test at the node. If x is continuous attributes, samples in D are need to be ordered. Assume that the ordered values are $w_1,\ldots,w_m$. Let $w=(w_i+w_{i+1})$, $i=1,2,\ldots, m-1$, according to which D can be split into two subsets, i.e. not greater than and greater than w. For each w, $gain_w$ is computed according to equation (5) and select the maximum as the local threshold T_2. Then D is split into two subsets

$$D_1 = \{w_j | w_j \leq T_2\},\ D_2 = \{w_j | w_j > T_2\} \tag{7}$$

Step 4: If D_i, i=1,2, is empty, the child node is set to be a leaf. If not empty, the DC approach is recursively applied to each child node until D satisfies a stopping criterion [20]. The stopping criterion can be that the samples in each node belong to one class or no attributes remain for partition.

If stopping criterion was tight, the decision tree created might be small and under-fitted. In contrary, if stopping criterion was loose, the decision tree built might be large and over-fitted [29]. To solve this dilema, pruning methods were developed. The C4.5 algorithm utilizes the pessimistic statistical correlation test to prune the tree. Let T denotes the tree built on D. The main idea is that the error ratio estimated using D is not reliable enough [12, 29]. Then continuity correction, which is a more realistic measure, is used

$$\varepsilon'(T,D) = \varepsilon(T,D) + \frac{|leaves(T)|}{2|D|} \tag{8}$$

where $\varepsilon(T,D)$ denotes the error rate of the tree T over the sample D and $|leaves(T)|$ indicates the number of leaves in T.

In equation 7, an optimistic error rate still produces. To solve this problem, an internal node t is pruned when its error rate is less than one standard error from a reference tree

$$\varepsilon'(pruned(T,t),D) = \varepsilon'(T,D) + \sqrt{\frac{\varepsilon'(T,D)(1-\varepsilon'(T,D))}{|D|}} \tag{9}$$

where $pruned(T,t)$ indicates the tree obtained by replacing the node t in T with a suitable leaf. If the classification error of a node is greater than the error of classifying all cases in D as belonging to the most frequent class in D, then the node is set to be a leaf and all subtrees are removed.

VII. Architecture of Classification Model

The architecture of the constructed model for PS classification is given in Fig. 3. It consists of seven parts: pulse signal acquisition system, pulse signal database (DB), class balance module, noise reduction module, baseline wander removal module, characteristic points detection module, parameter selection module and classification module based on DT. The pulse signal acquisition system, noise reduction module and baseline wander removal module were described in our previous work in detail [6]. The pulse signal acquisition system consists of preprocessing circuits and a pulse transducer. The preprocessing circuit is comprised of two amplifiers and an analog-digital converter. The pulse transducer is made by Shanghai University of Traditional Chinese Medicine. It is a duplex cantilever beam transducer, which can be distinguished from sensors that used in western medicine. The sensitivity and output impedance are 0.5millivolt per gramme (g) and one thousand ohm, respectively. This system can record a series of pulse signals under different contact pressures. The pulse signal whose modulus reaches the maximum is selected as the subject investigated [6].

In our study, the background noise and baseline wander are eliminated based on decomposition and reconstruction algorithm of wavelet, in which a smooth and symmetric wavelet function was used to the optimum decomposition scale is 4. The experiments showed that the method used to preprocess the pulse signal was effective [6].The characteristic points of pulse signals were detected based on complex-valued wavelet transform and were labeled based on chain code, which is testified that the proposed approach can detect the characteristics of the pulse signals accurately and is superior to the conventional techniques based on real-valued wavelet transform and the parameters can be estimated precisely. The detailed characteristic point detection procedure was illustrated in our studies [1, 6]. The ultimate goal of this model is to classify PS automatically. Firstly, pulse signal samples are collected and DB is balanced based on USMC algorithm. Secondly, pulse samples are preprocessed and the characteristic points are detected and labeled. Thirdly, the nine parameters are computed and selected. Fourthly, the obtained feature set is utilized as attributes of DT and the mapping relationships between the symptom set and the disease are constructed. Fifth, some classification rules can be extracted according to DT, which can be acquired from the root node to leaf nodes and the leaf nodes provide the predictive types. The rules can be employed to classify the unknown samples. Finally, the performance of the model is validated by experiments.

VIII. EXPERIMENTAL RESULTS

Our experiments will verify three objectives: (1) the proposed method can differentiate PS effectively; (2) the class balance module can achieve higher predictive accuracy rate (PAR); (3) the model incorporating parameters pre-selection procedure can improve PAR.

In DB, a total of over four hundreds pulse signals were gathered from several hospitals of TCM. The diagnostic results were given by one group of clinical physicians with three members. The sample number of R-pulse, F-pulse and NS-pulse are 198, 146 and 124, respectively.

In this work, a stratified κ-fold cross validation technique [21] (κ=10, named CV-10) is used to evaluate the model proposed. To explore the three objectives, we design the following experiments. Firstly, to eliminate the influence of class imbalance, we sample 124 cases based on USMC method for F-pulse and NS-pulse and build another DB, which is named B-DB. As comparison, we name the original DB as U-DB. Secondly, use MISS method to select parameters from U-DB and B-DB, respectively, and the survived parameter sets are expressed as V_1 and V_2. Thirdly, MBSS algorithm is employed to select parameters from U-DB and B-DB, respectively, and the obtained parameter sets are denoted by V_3 and V_4. Fourthly, use C4.5 algorithm to learn four DTs based on V_1, V_2, V_3, and V_4, respectively, and the models are denoted by D-M_1, D-M_2 D-M_3 and D-M_4. As comparison, the models built from U-DB and B-DB directly are expressed as D-M_5 and D-M_6, respectively.

After parameter selection module, $V_1=\{h_{sp}, h_{ff}, h_{ef}, r_{es}\}$, $V_2=\{h_{sp}, h_{ef}, h_{ff}, r_{es}, r_{fs}\}$, $V_3=V_4=\{h_{sp}, h_{ff}, r_{es}, r_{fs}, r_{fp}\}$. D-$M_1$ and D-M_2 obtain the same DTs. D-M_3 is more complex than D-M_4. In D-M_3, there are six leaves, while in D-M_4, five leaves remain. D-M_6 is more complex than D-M_5. D-M_6 has six leaves, while in D-M_5, there are five leaves. Fig. 4~6 sketch the DTs of D-M_2, D-M_4 and D-M_5. In D-M_2, only two parameters survive, i.e. h_{sp}, r_{es}. In D-M_4, three parameters survive, i.e. h_{sp}, r_{es}, r_{fp}. In D-M_5, three parameters survive, i.e. h_{sp}, r_{es}, h_{ef}. However, the three models all consist of five leaves, namely, they have the same complexity.

The confusion matrixes of the six models are shown as Table I~VI. Table VII lists the PAR comparison of these six models. The average PAR is above 90%, which testifies the efficiency and feasibility of our method. The PAR of D-M2 is higher than D-M1, D-M4 is higher than D-M3 and D-M6 is higher than D-M5. That validates the second objective.

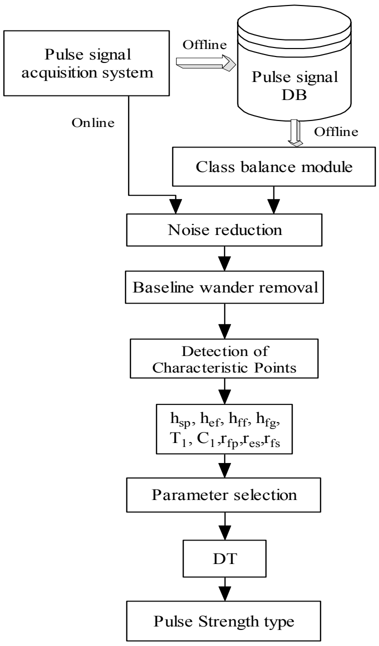


Figure 3. The architecture of PS classification model

It also can be noted that the influence of class balance module on the classification results is not very large in this work and the models built incorporating this module just obtain slightly better results than that without class balance procedure, which doesn't very accord with our first aim. This may be induced by several aspects. First, the classes in our DB are not very unbalanced and the number ratio between them is less than 2. Second, some useful information may be missed when sampling the database. Third, the scheme used for dealing with class imbalance may be not very suitable, which will be studied in our further work.

Some simple rules can be induced from DT. According to these rules, new pulse signal can be classified directly.

TABLE I.
CONFUSION MATRIX OF D-M1

	R-pulse	F-pulse	NS-pulse
R-pulse	185	0	13
F-pulse	0	136	10
NS-pulse	10	9	105

TABLE II.
CONFUSION MATRIX OF D-M2

	R-pulse	F-pulse	NS-pulse
R-pulse	115	0	9
F-pulse	0	116	8
NS-pulse	8	8	108

TABLE III.
CONFUSION MATRIX OF D-M3

	R-pulse	F-pulse	NS-pulse
R-pulse	187	2	9
F-pulse	6	134	6
NS-pulse	9	7	108

TABLE IV.
CONFUSION MATRIX OF D-M4

	R-pulse	F-pulse	NS-pulse
R-pulse	118	3	3
F-pulse	4	113	7
NS-pulse	7	5	112

TABLE V.
CONFUSION MATRIX OF D-M5

	R-pulse	F-pulse	NS-pulse
R-pulse	180	0	18
F-pulse	4	142	0
NS-pulse	18	6	100

TABLE VI.
CONFUSION MATRIX OF D-M6

	R-pulse	F-pulse	NS-pulse
R-pulse	106	0	18
F-pulse	0	124	0
NS-pulse	12	6	106

TABLE VII.
PAR COMPARISON

D-M1	0.9103±0.0448
D-M2	0.9113±0.0281
D-M3	0.9167±0.1912
D-M4	0.9220±0.1882
D-M5	0.9017±0.0261
D-M6	0.9032±0.0298

The rules induced from D-M$_5$ can be described as follows:

Rule 1: If $h_{sp} \le 14$, then the pulse signal is a F-pulse.

Rule 2: If $14 < h_{sp} \le 20.5$ and $h_{ef} \le 3.2$, then the pulse signal is a F-pulse.

Rule 3: If $14 < h_{sp} \le 20.5$ and $h_{ef} > 3.2$, then the pulse signal is a NS-pulse.

Rule 4: If $h_{sp} > 20.5$ and $r_{es} > 0.19$, then the pulse signal is a R-pulse.

Rule 5: If $h_{sp} > 20.5$ and $r_{es} \le 0.19$, then the pulse signal is a F-pulse.

The parameter h_{sp} is the height of percussion wave of pulse signal and h_{ef} is the height of tidal wave of pulse signal. The rules can be explained as following. Rule 1 shows that if percussion wave is small enough, the pulse signal is a F-pulse. Rule1 and rule 5 show that if percussion wave is not very small but tidal wave is small enough, the pulse signal is a F-pulse. Rule 3 denotes that if percussion wave and tidal wave are all moderate, the pulse signal is a NS-pulse. Rule 4 shows that if percussion wave and tidal wave are all large, the pulse signal is a R-pulse. These results correspond with the expertise of TCM, which was introduced in part III.

From Table VII, it can be seen that D-M$_1$, D-M$_2$ D-M$_3$ and D-M$_4$ get higher PAR than D-M$_5$ and D-M$_6$, which validate the third objective. The results also show that DT incorporating MBSS algorithm achieves slightly better results than MISS. This may be caused by that the parameters of pulse signal are not independent and they correlate with each other. MBSS algorithm uses BNs to select variables, which considers the dependent relationships among parameters adequately.

Rule1, rule 4 and rule 5 can also be deduced from D-M2 and D-M4. Based on D-M2, we can also write the following rules:

Rule 6: If $14 < h_{sp} \le 20.5$ and $r_{es} \le 0.19$, then the pulse signal is a F-pulse.

Rule 7: If $14 < h_{sp} \le 20.5$ and $r_{es} > 0.19$, then the pulse signal is a NS-pulse.

It can be seen that rule 5 and rule 6 can be merged and rewritten as following:

Rule 8: If $h_{sp} > 14$ and $r_{es} \le 0.19$, then the pulse signal is a F-pulse.

From D-M4, the following rules can be obtained:

Rule 9: If $14 < h_{sp} \le 20.5$ and $r_{fp} \le 0.182$, then the pulse signal is a F-pulse.

Rule 10: If $14 < h_{sp} \le 20.5$ and $r_{fp} > 0.182$, then the pulse signal is a NS-pulse.

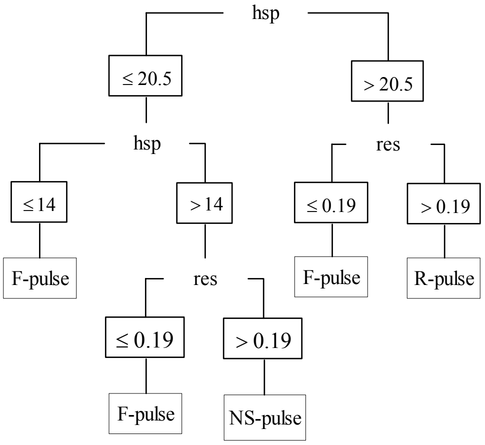


Figure 4. PS classification Model based on D-M2

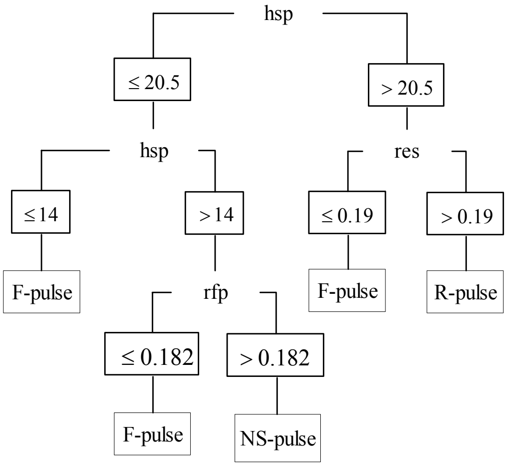


Figure 5. PS classification Model based on D-M4

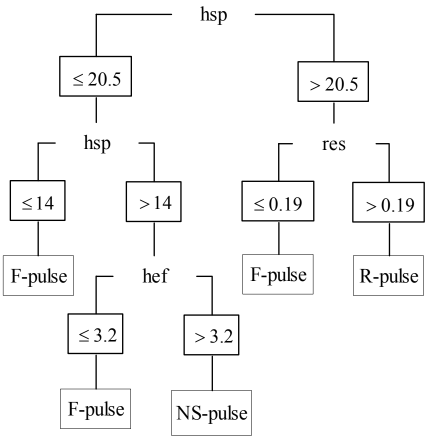


Figure 6. PS classification Model based on D-M5

The experiments testify that the proposed method is effective and can classify PS accurately. The six models all work well, which validate that the methodology we proposed is a good means to classify PS.

IX. Conclusion

Pulse diagnosis is one of the most important diagnostic methods in TCM, which can obtain important information of the health state of human body. But the complexity and elusiveness of traditional diagnostic method confine its development and generalization. Strength is one of important characteristics of pulse signal and its recognition is very complex. To explore an effective method for PS identification, we build classification models based on DT. The experiment results validate that the models incorporating parameters pre-selection procedure can achieve better results. The proposed method can classify PS accurately and provide an effective means for the analysis of pulse signal. This study is a beneficial exploration for quantification of pulse diagnosis.

Appendix A The Algorithm Description of Pulse Strength Classification Based on Decision Tree

A. Use under-sampling the majority class method to balance the pulse signal database.

B. Perform the preprocessing algorithm based on wavelet decomposition and reconstruction to remove noise and baseline wander of pulse signals.

C. Detect and label the characteristic points of pulse signals based on complex-valued wavelet and chain code.

D. Compute the parameters based on characteristic points.

E. Select the parameters based on BNs and get the symptom set.

F. Build the classification model based on the C4.5 algorithm.

G. Induce the classification rules based on the decision trees.

Acknowledgment

This work was supported in part by grants from the Scientific Research Fund of Zhejiang Provincial Education Department (Grant No. 20070593), the Major Program of Zhejiang Science and Technology Plan Item (Grant No. 2008c14100) and Zhejiang Natural Science Foundation of China (Grant No. Y1080230).

References

[1] H.Y. Wang, P.Y. Zhang, “Investigation on the automatic parameters extraction of pulse signals based on wavelet transform”, *Journal of Zhejiang University, Science A*, Vol. 8, No. 8, 2007, pp. 1283-1289.

[2] L. S. Xu, D. Zhang, K. Q. Wang, N. M. Li and X. Y. W, “Baseline wander correction in pulse waveforms using wavelet-based cascaded adaptive filter”, *Computers in Biology and Medicine*, Vol. 37, 2007, pp. 716-731.

[3] C.C. Chiu, S.J. Yeh, C.H. Chen, “Self-organizing arterial pressure pulse classification using neural networks: theoretical considerations and clinical applicability”. *Computers in Biology and Medicine*, Vol. 30, 2000, pp. 71-88.

[4] Hu J.N., Yan S.C., Wang X.Z., Chu H., 1997. “An intelligent Traditional Chinese Medicine pulse analysis system model based on artificial neural network”, *Journal of China Medical University*, Vol. 26, No. 2, pp. 134-137 (in Chinese).

[5] H.Y. Wang, P.Y. Zhang, “A model for automatic identification of human pulse signal”, *Journal of Zhejiang University, Science A*, Vol. 9, No. 10, 2008, pp. 1382-1389.

[6] P.Y. Zhang, , H.Y. Wang, “A framework for automatic time-domain characteristic parameters extraction of human pulse signals.” *Journal on Advances in Signal Processing*, Vol. 2008, Article ID 468390, 9 pages.

[7] Z.F. Fei, “Contemporary Sphygmology in Traditional Chinese Medicine”, People’s Medical Publishing House, Beijing, 2003. (in Chinese).

[8] M. Pal, P.M. Mather, “Decision tree based classification of remotely sensed data”, *Proceedings of the 22nd Asian Conference of Remote Sensing*, Singapore, November, 2001, pp. 5-9.

[9] J.R. Quinlan, “Induction of Decision Tree”, *Machine Learning*, Vol. 1, No. 1, 1996, pp. 81-106.

[10] J. Cheng, U.M. Fayyad, K.B. Irani and Z. Qian, “Improved decision trees: a generalized version of ID3”, *Proc. 5th Int. Conf. on Machine Learning*, Morgan Kaufmann, San Mateo, CA, 1988, pp. 100-108.

[11] U.M. Fayyad, K.B. Irani, “A machine learning algorithm (GID3*) for automated knowledge acquisition: improvements and extensions,” GM Research labs, 1991, Warren MI.

[12] J.R. Quinlan, “C4.5: Programming for machine learning”, San Mateo, Calif. Morgan Kaufmann, 1993.

[13] S. Ruggieri, “Efficient C4.5”, *IEEE Transaction on Knowledge and Data Engineering*, Vol. 14, No. 2, 2002, pp. 438-444.

[14] B. H. Jun, C. S. Kim, H. Y. Song, J. Kim, “A new criterion in selection and discretization of attributes for the generation of decision tree”, *IEEE Transaction on Pattern Analysis and Machine Intelligence*, Vol. 19, No. 12, 1997, pp. 1371-1375.

[15] G. M WEISS, “Mining with rarity：A unifying framework”, Chicago, IL, USA, *SIGKDD Exploration*, 2004, Vol. 6, No. 1, pp. 7-19.

[16] R. Barandela, J.S., Sanchez, V. Garcia, E. Rangel, “Strategies for learning in class imbalance problems”, *Pattern Recognition*, Vol. 36, No. 3, 2003, pp. 849-851.

[17] V. C. Nitesh, “Smote: synthetic minority over-sampling technique”, *Journal of Artificial Intelligence Research*, Vol. 16, 2002, pp. 321-357.

[18] M. Kubat, S. Matwin, “Addressing the curse of imbalanced training sets: one-sided selection”, *Proc. 14th Int'l Conf. Machine Learning*, 1997, pp. 179-186.

[19] D.S. Wu, “Detecting information technology impact on firm performance using DEA and decision tree”, *International Journal of Information Technology and Management*, Vol. 5, No. 2-3, 2006, pp. 162-174.

[20] J.R. Quinlan, “Improved use of continuous attributes in C4.5”, *Journal of Artificial Intelligence Research*, Vol. 4, 1996, pp. 77-90.

[21] M. Mullin and R. Sukthankar, “Complete cross-validation for nearest neighbor classifiers,” *Proceedings of ICML*, 2000, pp. 639-646.

[22] H.Y. Wang, J. Wang, “A Quantitative Diagnostic Method Based on Bayesian Networks in Traditional Chinese Medicine”, *Lecture notes in computer science*, Vol. 4234, 2006, pp.176-183.

[23] H.Y. Wang, “A Computerized Diagnostic Model Based on Naive Bayesian Classifier in Traditional Chinese Medicine”, *Proceedings of the First International Conference on Biomedical Engineering and Informatics*, Vol.1, 2008, pp.474-477

[24] G. Isabelle, “An Introduction to Variable and Feature Selection”, *Journal of Machine Learning Research*, Vol. 3, pp.1157-1182.

[25] R. Kohavi, G. John, “Wrappers for Feature Selection”, *Artificial Intelligence*, Vol. 97, No. 1, 1997, pp.273-324.

[26] A.K.C. Wong, D.K.Y. Chiu, “Synthesizing statistical knowledge from incomplete mixed-mode data”, *IEEE Trans. Pattern Anal. Mach. Intell.* Vol. 9, No. 6, 1987, pp.796-805.

[27] P. Spirtes, C. Meek, “Learning Bayesian networks with discrete variables from data”, *Proceedings of the First International Conference on Knowledge Discovery and Data Mining*, Menlo Park, CA: AAAI Press, 1995, pp.294-299.

[28] R.E. Neapolitan, “Probabilistic Reasoning in Expert Systems：Theory and Algorithms”, John Wiley and Sons, New York, NY, 1990.

[29] L. Rokach, O. Maimon, “Top-down induction of decision trees classifier- A survey”, *IEEE Transaction on System, Man, and Cybernetics*, Vol. 35, No. 4, pp.476-487.

Huiyan Wang was born in Yantai, Shandong province, China, on April 3, 1975. She received her Ph.D. degree in electric engineering from Zhejiang University, China, in 2004.

She was a postdoctoral researcher on quantitative diagnosis in TCM at pharmaceutical information institute of Zhejiang University, from 2003 to 2005. She is currently an associate professor with the college of Computer Science and Information Engineering, Zhejiang Gongshang University. Her research interests are signal processing, pattern recognition and data mining.

Peiyong Zhang was born in Anqing, Anhui province, China, on May 18, in 1977. He received his Ph.D. degree from Zhejiang University in 2004.

He is currently an associate professor with the Institute of VLSI Design, Zhejiang University. His research interests are VLSI Design for Manufacturability, signal processing.