Intelligent Categorization Method for Diagnosing Cardiovascular Diseases Hierarchically

Mubo Chen and Mingchui Dong
Dept. of ECE, FST, University of Macau, Macau S. A. R., China
Email: chenmubocs@gmail.com

***Abstract*—For detecting cardiovascular diseases (CVDs) hierarchically, it has been verified that a disease-oriented categorization method is necessary and valid. Such method should be able to categorize 32 hemodynamic parameters (HDPs), 9 symptom parameters (SPs), and 6 physiological parameters (PPs) into different groups aiming at diagnosing different CVDs efficiently. To this end, IGAECM, a new categorization method based on information gain attribute evaluation with two pivotal steps is proposed. First, compute the information gain (IG) of each HDP, SP or PP variable, and discard redundant variables with zero IG. Then according to IG of each variable, search and categorize the remaining HDP, SP and PP variables into different groups automatically. Compared with previous proposed method CARTCM (categorization method based on classification and regression tree), it has three main advantages: (i) an IG-based searching strategy is proposed to group HDPs, SPs & PPs while CARTCM uses a simple threshold strategy; (ii) automatically determines not only the number of groups but also the quantity and types of parameters relevant to CVDin each group, while CARTCM contains user manually defined thresholds; (iii) discards the redundant variables thus reduces the computing complexity while CARTCM does not. The effectiveness and adaptability of such method is demonstrated and tested by applying it in diagnosing 5 most common and important CVDs successfully.**

***Index Terms*—cardiovascular diseases, disease-oriented, information gain based searching strategy, multi-label learning**

I. Introduction

To date, cardiovascular diseases (CVDs) have emerged as the top health killer in both urban and rural areas in most of the countries. There among, coronary heart disease (CHD), hypertension (HT), hyperlipemia (HL), arrhythmia (AR), and cerebral infarction (CIN) are five frequently encountered and typical CVDs harming people's health [1]. Due to this reason, CVDs detection has attracted more and more interests from medical researchers and data scientists [2]-[6].

CVDs can be detected by diagnosing vital signs such as electrocardiography (ECG) [4], echocardiography [7], heart sound (HS) [8] and/or sphygmogram (SPG) [9]. Among them, SPG is competitive and widely adopted in e-home healthcare usage for its non-invasive and easy acquisition. Based on SPG signals, hemodynamic parameters (HDPs) are derived by using the model of elastic cavity, which are capable of revealing cardiovascular health status and variation tendency [9]. Besides, in medical theory, a symptom is expression of the presence of disease or abnormality. Hence, hard efforts of exploring HDPs, symptom parameters (SPs) as well as physiological parameters (PPs) have been made for CVDs detection by applying machine learning technology [10]-[13]. There exist many models for exploring HDPs, SPs and PPs, but the most important distinction is: does the model put the parameters into a classifier all at once, or does the model put the parameters hierarchically. For the former case, Ref. [11] combines association analysis and information gain feature selection for SPs on multi-syndrome data of CHD considering the association of SPs; Ref. [12] proposes a hybrid optimization based on multi-label feature selection to effectively reduce the data dimension and improves the classification performance on CHD; Ref. [13] shows high accuracy in detecting CVDs by using SVM (support vector machine) based on HDPs and PPs, etc. However, the aforementioned methods may contradict to the doctors' clinical diagnosis procedure. In practice, doctor normally ranks all parameters and selects specified ones with most pertinence to diagnose diseases. If it fails, doctor would turn to the less pertinent parameters. Such a manner makes the clinical reasoning procedure representing "hierarchically" character. For this reason, detecting CVDs hierarchically by using machine learning has attracts more and more attention. However, the hierarchical mode inevitably leads to a bottleneck problem: How to divide HDPs, SPs& PPsinto groups specifically relevant to different CVDs so as to construct a hierarchical classifier?

At present, researchers have made several attempts to tackle such a formidable problem. A generic HDP&PP categorization method based on one-way analysis of variance (ANOVA) is adopted in [14] before constructing hierarchical fuzzy neural networks (HFNNs) [15]. As a result, HDPs and PPs are categorized into sensitive, supporting, inertia groups. However, such a generic HDP&PP categorization is improper or even invalid for detecting CVDs due to the fact that HDP&SP&PP categorization should be specific to different CVDs. Thus, Ref. [16] takes into account the diversity of HDP&PP categorization and proposes a categorization method based on classification and regression tree (CARTCM),

Manuscript received August 21, 2014; revised November 1, 2014.

doi: 10.12720/joace.3.6.512-518

which shows better classification performance. Nevertheless, the user manually setting thresholds adopted in CARTCM limits its robustness and stability in practical application. Alternatively, this research proposes a more intelligent categorization method IGAECM, which is based on information gain attribute evaluation technique. Instead of manually setting thresholds, in IGAECM, an IG-based searching strategy is adopted to categorize automatically the HDP&SP&PP variables into groups. Its theoretical explanation is presented, the effectiveness and validness of IGAECM are verified by experiments testing the real data set sampled from patients with diseases CHD/HT/HL/AR/CIN.

TABLE I. PARTIAL RECORDS OF SITE-MEASURED CVD RELEVANT DATA

ID	Name	Age	Dizzy	…	MST	…	CIN
1	Patient 1	80	0	…	20.93	…	L_{15}
2	Patient 2	49	1	…	17.82	…	L_{15}
…	…	…	…	…	…	…	…

II. METHODS

A. Pre-Processing of Site-Measured CVD Relevant Data

As shown in Table I, the site-measured medical records consist of patients' personal information, PPs, SPs and HDPs as well as doctors' diagnostic results. Denote each HDP&SP&PP record as $\mathbf{x}_i = [x_{1i}, x_{2i}, \ldots, x_{Ni}]^{\mathrm{T}}$ ($i = 1, 2, \ldots, M$), where M is the total number of patients' records and N is the total number of HDP&SP&PP. Specifically, x_{ji} is the i^{th} record's j^{th} HDP&SP&PP, where $j = 1, 2, \ldots, N$. Thus denote all patients'HDP&SP&PP records as matrix $\mathbf{X}_{N \times M}$:

$$\mathbf{X}_{N\times M}=[\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_M] \quad (1)$$

The diagnostic result of the i^{th} record is denoted as: $\mathbf{y}_i = [y_{1i}, y_{2i}, \ldots, y_{Di}]^{\mathrm{T}}$ ($i = 1, 2, \ldots, M$), where D is the total number of CVD types and specifically $y_{di} \in \{L_{1d}, L_{2d}\}$is diagnostic result of the i^{th} record's d^{th} CVD.$y_{di}=L_{1d}$ represents that the patient with the i^{th} record does not have the d^{th} CVD, while $y_{di}= L_{2d}$means that the patient with the i^{th} record has the d^{th}CVD. Thus, $\mathbf{Y}_{D \times M}$ obtained by medical inference is denoted as:

$$\mathbf{Y}_{D \times M} = [\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_M] \quad (2)$$

In this research, totally 47 parameters including 32 HDPs, 9 SPs and 6 PPs (the acronyms of HDPs, SPs &PPs can be found in Table II) are employed to diagnose five types of typical and frequently encountered CVDs (CHD, HT, HL, AR, CIN), thus N=47 and D=5.

The diagnostic aim is to infer the diagnostic result $\mathbf{Y}_{D \times M}$ from site-measured $\mathbf{X}_{N \times M}$ as:

$$\mathbf{Y}_{D\times M} = \mathbf{S}_{D\times N} \bullet \mathbf{X}_{N\times M} \quad (3)$$

where $\mathbf{S}_{D \times N}$ is a nonlinear mapping which can be realized by SVM, HFNNs or so on. Obviously, using all 47 HDPs, SPs and PPs at once to detect CVDs is paradoxical, influencing diagnostic accuracy, high computation-consuming and unconformable to doctors' normal clinical diagnosis procedure. Thus, HDP&SP&PP categorization is necessary which is further depicted below.

TABLE II. ACRONYMS LIST OF HDPS, SPS, PPS, AND CVDS

HDP&SP&PP ACRONYM	EXPLANATION
CHDH	CHD history
HTH	HT history
HLH	HL history
CP	Chest pain
SK	Smoking
DK	Drinking
AG	Age
SX	Sex
CD	Chest distress
PAL	Palpitation
DY	Dizzy
HEI	Height
WEI	Weight
SP	Systolic pressure
DP	Diastolic pressure
PR	Pulse rate
SI	Stroke index
SV	Stroke volume
CO	Cardiac output
CI	Cardiac index
VPK	Ventricular power coefficient
EMK	Left myocardial power utilization coefficient
SWI	Left ventricular power utilization coefficient
HOI	Heart oxygen index
HOV	Heart oxygen volume
CMBV	Cardiac muscle blood volume
CMBR	Cardiac muscle blood supplying rate
PP	Pulse pressure
MSP	Mean systolic pressure
MDP	Mean diastolic pressure
MAP	Mean arterial pressure
CCP	Coronary artery perusing pressure
BEK	Blood vessel evacuating coefficient
AC	Artery compliance
FEK	Flexibility ectasia coefficient
TPR	Total peripheral resistance
SPR	Standard peripheral resistance
VER	Left ventricle eject resistance
PAMP	Pulmonary arterial wedge pressure
PAR	Pulmonary arterial resistance
PAP	Pulmonary arterial pressure
BV	Blood volume
TBV	Total blood viscosity
TBVR	Total blood viscosity restored
MHR	Microcirculation half refresh ratio
MRT	Microcirculation half refresh time
MST	Microcirculation mean stagnation time
CHD	Coronary heart disease
HT	Hypertension disease
HL	Hyperlipemia disease
AR	Arrhythmia disease
CIN	Cerebral infarction disease

B. IGAECM

The IGAECM shown in Fig. 1 includes two parts: 1) compute the IG of each HDP&SP&PP variable; 2) categorize automatically HDPs, SPs & PPs into *gd* groups using an IG-based searching strategy (IGSS) according to the pre-detected CVD, and correspondingly train *gd* classifiers whose classification accuracies are declined from Classifier_{1d} to Classifier_{gd}, where $d = 1, 2, \ldots, D$ and g is the total group number for the d^{th} CVD.

PART I: Compute IG for each variable

Diverse feature selection techniques have been proposed to discard redundant and irrelevant features, and the following attribute evaluations are frequently used: IG [17], gain ratio [18], symmetrical uncertainty [19], etc. In this research, IG attribute evaluation (IGAE) is adopted.

As the foundation of IGAE, entropy is the measure of indeterminacy for random variable. Let $L_d \in \{L_{1d}, L_{2d}\}$ be diagnostic result and A_{jd} (j = 1, 2, … , N; d = 1, 2, … , D) be selected the j^{th} HDP&SP&PP variable for detecting the d^{th} CVD. Consequently, the entropy of classification system can be defined as

$$H(L_d) = -\sum_{l\in\{1,2\}} P(L_{ld})\log_2\left(P(L_{ld})\right) \quad (4)$$

And the IG of A_{jd} can be defined as

$$IG(A_{jd}) = H(L_d) - H(L_d|A_{jd})$$

$$= -\sum_{l\in\{1,2\}} P(L_{ld})\log_2\left(P(L_{ld})\right) + \sum_{f}^{F} P(A_{jdf}) \sum_{l\in\{1,2\}} P(L_{ld}|A_{jdf})\log_2(P(L_{ld}|A_{jdf})) \quad (5)$$

where F is the total number of values for A_{jd};A_{jdf} is the f^{th} value for j^{th} HDP&SP&PP variable aiming at the d^{th} CVD; $P(L_{ld})$ is the marginal probability of L_{ld} for the d^{th} CVD; $P(A_{jdf})$ is the marginal probability for A_{jdf}; and $P(L_{ld}|A_{jdf})$ is the conditional probability of L_{ld} given A_{jdf}. Then IG of each HDP, SP or PP variable is calculated respectively. As a result, variables with zero IG are distinguished as redundant features and discarded since they make no contribution to diagnosing CVDs.

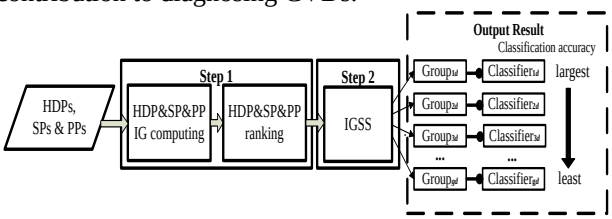

Figure 1. The construction of IGAECM.

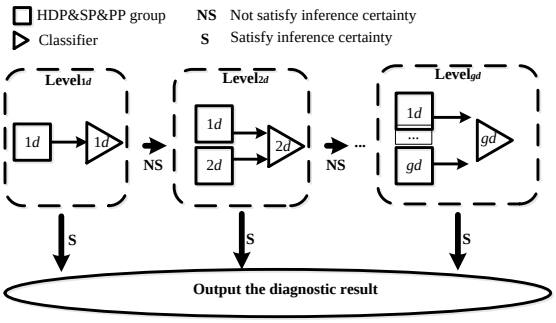

Figure 2. Schematic depiction of hierarchical classifier.

PART II: IGSS

In practice, HDPs, SPs & PPs with lager IG show high relevance and thus should be selected firstly to diagnose diseases. Naturedly, the next thing to do is to rank the remaining variables with their IGs and the ranked variables focusing on the d^{th} CVD are denoted as A'_{jd}(j = 1, 2, … , B_d, and B_d is the total number of remaining variables for the d^{th} CVD). In hierarchical diagnostic system [20] shown in Fig. 2, 47 HDPs, SPs & PPs are categorized into varied groups aiming at diagnosing different types of CVDs. For detecting the d^{th} CVD, firstly, variables in Group_{1d} are selected to feed into classifier_{1d}. Consequently, the classifier outputs the diagnostic result with probability estimate. If its probability satisfies the inference certainty (larger than the user-defined threshold θ_{1d}), then the inference stops and obtains the final diagnostic result, which of course is the best case. Otherwise, variables in next group combined with previous group (Group_{1d}) will be chosen to feed into classifier_{2d}. The same inference mechanism is repeated until Group_{gd}. In the worst case, all 47 HDPs, SPs& PPs (variables in Group_{1d}, Group_{2d}, … , Group_{gd} are merged) are used and fed to Classifier_{gd}, thus outputs the diagnostic result. Obviously, such a hierarchical mechanism should obey two rules:

Rule 1: The classifier_{1d} should be fed with the most correlated HDP&SP&PP group. On one hand, this rule is consistent with doctor's diagnostic procedure which selects the most specific and correlated parameters to detect CVDs. On the other hand, classifier_{1d} should be able to classify correctly a large proportion of HDP&SP&PP records with high certainty, which will speed up the diagnostic procedure.

Rule 2: The classification accuracy of $\text{classifier}_{(r+1)d}$ should be larger than that of classifier $_{rd}$. This rule makes sure that classifier $_{(r+1)d}$ is more likely to classify correctly the difficult and stubborn HDP&SP&SP records which could not be classified by classifier $_{rd}$.

The big question is, how to categorize HDPs, SPs& PPs into groups so as to construct a hierarchical classifier obeying the above rules? In this research, an IGSS integrated with hill-climb and SVM is proposed to solve this intractable problem. There among, hill-climbing is a greedy mathematical optimization technique which belongs to the family of local search [21] and is adopted here to select HDP&SP&PP variables and put them to the optimal groups; SVM [22] is used to be the classifier that evaluates the performance (accuracy) of the classification using hill-climb. Details of IGAECM with IG computing and IGSS are shown below.

Algorithm:

Input: HDP&SP&PP variables A_{jd} (j = 1, 2, … , N;d = 1, 2, … , D) and training set of HDP&SP&PP records.

Output: HDP&SP&PP categorizations aiming at different types of CVDs.

For d = 1, 2, … , D

(a)Compute IG of each HDP, SP or PP variable specific to the d^{th} CVD and discard those variables with zero IG.

(b) Rank the remaining variables with their IGs and then get A'_{jd} (j = 1, 2, … , B_d) where their IGs are ranged from the largest to the least.

(c) Set k = 1, q = 1.

(d) Put A'_{kd} into Group'_{qd}, and train the SVM classifier with Group'_{qd} where the classifier accuracy is recorded as Acc_k.

(e) Let $k = k + 1$, $k' = k$.Repeat step (d) and get the new Acc_k. If $Acc_k \geqslant Acc_{k'}$, repeat step (e); otherwise, Group'_{qd} will be the new group, and the classifier

accuracy $GroupAcc_{qd}$ of this group will be denoted as $Acc_{k'}$. Afterwards, let $q = q + 1$, and repeat step (d). The loop is running until $k = B_d$.

(f) Rank the groups with their corresponding classifier accuracies ranged from the largest to the least. If two or more than two classifier accuracies are same, then merge the corresponding groups into one, retrain the merged group and again compare it with the other remaining groups.

(g) Finally generate the ranked HDP&SP&PP groups which are presented as $Group_{qd}$, where $d = 1, 2, \dots, D$;$q = 1, 2, \dots, g$; and g is the number of groups relevant to the d^{th} CVD. It is easy to find that the hierarchical system with such generated groups satisfies both Rules 1 and 2. For Rule 1, $Group_{1d}$ is the most relevant categorized group. While for Rule 2, $Classifer_{2d}$ by using $Group_{1d}$ and $Group_{2d}$ is obviously stronger than $Classifer_{1d}$utilizing $Group_{1d}$.

III. Test Results

356 patients' records and 151 non-patients' records including 346 CHD, 257 HT, 34 HL, 157 AR, 37 CIN records (partially shown in Table I) obtained from Beijing Changping Chinese Medicine Hospital and the Second Affiliated Hospital of GuangXi Medical University are selected as testing samples. Some patients may get more than one types of CVDs. With random selection, 304 records (60%) and 203 records (40%) are selected as training set and testing set, respectively. The HDP&SP&PP categorization results of IGAECM and CARTCM are shown in Table III.

TABLE III. Disease-Oriented HDP&SP&PP Categorization

Diseases	Group	CARTCM	IGAECM
CHD	$Group_{11}$	AG, PAWP, PAP, EMK, FEK, DP	CD, SI, HTH, EMK, PAL, VER, VPK, HOV, PAP, PAWP, FEK, SPR, BEK, CMBR, SV, CMBV, SV, CP, PP, SWI, HOI, MRT, MST, MHR, AC, PR, DP, TPR, MSP, DY, CI, CO, CCP, SK, HEI, DK
	$Group_{21}$	CHDH, CMBR, HOV, SV, WEI, CO, HOI, PR, CMBV, CP, SI, SWI, CCP, VPK, HTH, CI, SP, AC, HEI	AG, SP, CHDH
	$Group_{31}$	HLH, SK, DK, SX, CD, PAL, DY, PP, MSP, MDP, MAP, BEK, TPR, SPR, VER, PAR, BV, TBV, TBVR, MHR, MRT, MST	WEI, SX, HLH
HT	$Group_{12}$	HTH, AG, SP, VPK, PR	SI, HEI, HOV, MAP, CP, MSP, CHDH, SV, SPR, DP, CD, CMBV, FEK, PAL, SK, HLH, SX, DK, DY
	$Group_{22}$	VER, CP, EMK, SPR, MSP, TPR, MDP, MAP, PP, SV, DP, PAR, HOV, TBV, BV, CI, PAL, HOI, BEK, WEI, CO, HEI, SI, CHDH	AG, HTH, SP, PR, EMK, VPK, PP
	$Group_{32}$	HLH, SK, DK, SX, CD, DY, SWI, CMBV, CMBR, CCP, AC, FEK, PAWP, PAP, TBVR, MHR, MRT, MST	VER, SWI
HL	$Group_{13}$	AG, PR, HOI, HOV	AG, PAP, PAWP, EMK, VER, HTH, CD, DY, SK, DK, CP, CHDH, SX, HLH, PAL
	$Group_{23}$	TBV, VER, EMK, SPR, CP, CCP, BV, BEK, TPR, DP, MRT, MST, CHDH, TBVR, PP, MHR	N/A
	$Group_{33}$	HTH, HLH, SK, DK, SX, CD, PAL, DY, HEI, WEI, SP, SI, SV, CO, CI, VPK, SWI, CMBV, CMBR, MSP, MDP, MAP, AC, FEK, PAWP, PAR, PAP	N/A
AR	$Group_{14}$	AG, CP, HTH, MAP, VPK, SP, MSP, MDP, SWI, CO, WEI	PAR, CD, HLH, SX, DY, CHDH, PAL, SK, DK
	$Group_{24}$	SV, DP, DY, CCP, CMBV, TPR, CI, PAR, FEK, EMK, CMBR, PP, HOI, TBV, BV, AC, VER, PAWP, PAP, CHDH, HEI, PR, BEK, CD, SI, DK, SPR, HOV, MRT, MST	CP, VPK, AG, SPR, HTH
	$Group_{34}$	HLH, SK, SX, PAL, TBVR, MHR	PP, SV, SI, CMBV, MAP, MSP, PR
	$Group_{44}$	N/A	SP, WEI, MDP, SWI, DP
CIN	$Group_{15}$	AG, PR, SP, MAP, VPK, CP, SV, SPR, TPR, MSP, DY, WEI	AG, SP, HTH, DY, PP, PR, VPK, TBV, HOV, CHDH, SK, CP, CD
	$Group_{25}$	HTH, PAR, VER, TBVR, TBV, HOI, SI, HEI, MDP, CMBV, CCP, CI, PP, BEK, CD, SWI, MHR, PAL, CO, CHDH	PAL, SX
	$Group_{35}$	HLH, SK, DK, SX, DP, EMK, HOV, CMBR, AC, FEK, PAWP, PAP, BV, MRT, MST	HLH, DK

In IGAECM, HDP&SP&PP parameters are categorized into 3, 3, 1, 4, 3 groups in terms of CHD, HT, HL, AR, CIN, respectively. Specifically, for CHD, 36 parameters are selected as $Group_{11}$, 3 for $Group_{21}$, and 3 for $Group_{31}$; for HT, 19 parameters are selected as $Group_{12}$, 7 for $Group_{22}$, and 2 for $Group_{32}$; for HL, 15 parameters are selected as $Group_{13}$; for AR, 9 parameters are selected as $Group_{14}$, 5 for $Group_{24}$, 7 for $Group_{34}$, and

5 for $Group_{44}$; for CIN, 13 parameters are selected as $Group_{15}$, 2 for $Group_{25}$, and 2 for $Group_{35}$.

TABLE IV. ACCURACIES OF CLASSIFIERS USING GROUPS WITHOUT HIERARCHY

Classifer	CHD (%)	HT (%)	HL (%)	AR (%)	CIN (%)
$Group_{1d}$	98.68	95.07	98.68	86.84	99.34
$Group_{2d}$	93.42	92.43	N/A	85.20	82.56
$Group_{3d}$	68.09	49.67	N/A	70.72	82.27
$Group_{4d}$	N/A	N/A	N/A	70.07	N/A

Table IV shows the accuracies of non-hierarchical classifiers using IGAECM to categorize HDP&SP&PP groups in training. LIBSVM [23] with prevalent C-support vector classification model and Radial Basis Functional kernel are used to construct such classifiers. It is easy to find that the diagnostic performance using $Group_{1d}$ is better than that using $Group_{2d}$ of the category, and so forth, which meets the aforementioned Rule 1 of designing a hierarchical diagnostic system. While for CARTCM, total HDP&SP&PP parameters are categorized into 3 groups in terms of all diseases. Specifically, for CHD, 6 parameters are selected as $Group_{11}$, 19 for $Group_{21}$, and 22 for $Group_{31}$; for HT, 5 parameters are selected as $Group_{21}$, 24 for $Group_{22}$, and 18 for $Group_{32}$; for HL, 4 parameters are selected as $Group_{13}$, 16 for $Group_{23}$, and 27 for $Group_{33}$; for AR, 11 parameters are selected as $Group_{14}$, 30 for $Group_{24}$, 6 for $Group_{34}$; for CIN, 12 parameters are selected as $Group_{15}$, 20 for $Group_{25}$, and 15 for $Group_{35}$. The reason why the categorization results of CARTCM and IGAECM are different is three-folds. Firstly, IGAECM discards the redundant variables with zero IG, while CARTCM reserves all the variables. Secondly, the attribute evaluations are different: IGAECM adopted IG as attribute evaluation, while Gini coefficient is used in CARTCM. Finally, IGAECM proposes an IG-based search strategy which is more intelligent and effective to solve the aforementioned formidable categorization problem, while CARTCM adopts only a simple user manually setting thresholds strategy.

TABLE V. ACCURACIES OF CLASSIFIERS USING GROUPS WITH HIERARCHY

Classifer	CHD (%)	HT (%)	HL (%)	AR (%)	CIN (%)
$Classifer_{1d}$	98.68	95.07	98.68	86.84	99.34
$Classifer_{2d}$	99.67	100	N/A	99.01	99.34
$Classifer_{3d}$	99.67	100	N/A	99.67	99.34
$Classifer_{4d}$	N/A	N/A	N/A	100	N/A

In order to verify the effectiveness of categorization using IGAECM, a hierarchical classifier shown in Fig. 2 are constructed. From Table V, it is easy to find that higher level classifiers with categorized groups might gain better classification performance than the lower level ones. Take diagnosing CVD as an example, the accuracy of $Classifier_{21}$ with $Group_{11}$ and $Group_{21}$ is better than that of $Classifier_{11}$ with only $Group_{11}$, and so on. That is to say, our proposed categorization method also satisfies the aforesaid Rule 2. Thenceforth such a hierarchical classifier is applied to detect CVDs on the unknown testing dataset. Here, multi-label evaluation metrics proposed in [24] are used in this research:

1) Hamming loss

Evaluates how many times an HDP&SP&PP variables-CVDs pair is misclassified, i.e., one CVD not relevant to the HDP&SP&PP record is predicted or one CVD relevant to the record is not predicted. The performance is perfect when hamming loss hloss(h) = 0 where h is the hierarchical classifier; the smaller the value of hloss(h), the better the performance.

$$\text{hloss}(h)=\frac{1}{|C|}\sum_{c=1}^{|C|}\frac{1}{D}|h(\mathbf{x}_c)\Delta\mathbf{y}_c| \quad (6)$$

where $\triangle$ stands for the symmetric difference between two sets, $|C|$ is the total number of testing set, D is the total number of CVDs, $(\mathbf{x}_c, \mathbf{y}_c)$ is the c^{th} HDP&SP&PP variables-CVDs pair in testing set.

In detecting CVDs, the hierarchical classifier will not only produce what types of CVDs the patient may have, but also generate a real-valued function $f(.,.)$ to depict the probability of having diseases. A successful learning classifier tends to output larger $f(.,.)$ for diseases in $\mathbf{y}_c$, i.e. $f(\mathbf{x}_c, y_1) > f(\mathbf{x}_c, y_2)$ for $y_1 \in \mathbf{y}_c$ and $y_2 \notin \mathbf{y}_c$. Hence the following evaluations based on $f(.,.)$ are concerned:

2) One-error

Evaluates how many times the top-ranked predicted disease is not in the set of proper diseases for the HDP&SP&PP record. The performance is perfect when one-error(f) = 0; the smaller the value of one-error(f), the better the performance.

$$\text{one-error}(f)=\frac{1}{|C|}\sum_{c=1}^{|C|}\left[\left[\left[\arg\max_{y\in Y'} f(\mathbf{x}_c, y)\right]\notin\mathbf{y}_c\right]\right] \quad (7)$$

where for any predicated π, $[\![\pi]\!]$ equals 1 if π holds and 0 otherwise, and Y' is the set of {CHD, HT, HL, AR, CIN} in this research.

3) Ranking loss

Evaluates the average fraction of predicted disease pairs that are reversely ordered for the HDP&SP&PP record. The performance is perfect when rloss(f) = 0; the smaller the value of rloss(f), the better the performance.

$$\text{rloss}(f)=\frac{1}{|C|}\sum_{c=1}^{|C|}\frac{1}{|\mathbf{y}_c||\bar{\mathbf{y}}_c|}|\{(y_1, y_2)|f(x_c, y_1)\le f(\mathbf{x}_c, y_2)| \quad (8)$$

where $(y_1, y_2)\in\{\mathbf{y}_c\times\bar{\mathbf{y}}_c\}$, $\bar{\mathbf{y}}_c$ is the complementary set of $\mathbf{y}_c$ in Y' and $|\mathbf{y}_c|$ is the number of predicted diseases.

4) Coverage

Evaluates how far, on the average, to go down the list of predicted diseases in order to cover all proper diseases of the HDP&SP&PP record. The smaller the value of coverage(f), the better the performance.

$$\text{coverage}(f)=\frac{1}{|C|}\sum_{c=1}^{|C|}\max_{y\in\mathbf{y}_c} rank(\mathbf{x}_c, y)-1 \quad (9)$$

where rank(., .) is the function that maps the outputs of $f(\mathbf{x}_c, y)$ for any $y\in Y'$ to {1, 2, … , D} such that if $f(\mathbf{x}_c, y_1) > f(\mathbf{x}_c, y_2)$ then $rank(\mathbf{x}_c, y_1) < rank(\mathbf{x}_c, y_2)$.

5) Average precision

Evaluates the average fraction of predicted diseases ranked above a particular label $y \in Y'$ which actually are in Y'. The performance is perfect when avgprec(f) = 1; the bigger the value of avgprec(f), the better the performance.

$$\text{avgprec}(f)= \frac{1}{|C|}\sum_{c=1}^{|C|}\frac{1}{|\mathbf{y}_c|}\sum_{y\in \mathbf{y}_c}\frac{|\{y'|rank(\mathbf{x}_c,y')\le rank(\mathbf{x}_c,y), y'\in \mathbf{y}_c\}|}{rank(\mathbf{x}_c,y)} \quad (10)$$

Equipped with the above evaluation metrics, the prediction accuracies of hamming loss, one-error, coverage, average precision for the hierarchical classifier by utilizing IGAECM are shown in Table VI comparing with that using CARTCM. In this research, the probability threshold θ of each level is set uniformly to 0.9 so as to output diagnostic result with high certainty. The values of the five evaluation metrics corresponding to IGAECM are as follows: hamming loss 0.2591, 0.0079 lower than the CARCM result 0.2670; coverage 0.9704, 0.0887 lower than CARCM 1.0591; ranking loss 0.0610, 0.0264 lower than CARTCM 0.0874; average precision 0.9647, 0.0061 larger than CARTCM 0.9586; while one-error 0.0211, 0.0141 larger than CARTCM 0.0070. The above results show that IGAECM has better performance in hamming loss, coverage, ranking loss, average precision than CARTCM while gets fair performance in one-error which need further improvement.

TABLE VI. COMPARISON OF CARTCM AND IGAECM ON CVD RELEVANT TESTING DATASET

Evaluation metric	CARTCM	IGAECM	Comparison
hamming loss	0.2670	0.2591	↓ 0.0079
coverage	1.0591	0.9704	↓ 0.0887
ranking loss	0.0874	0.0610	↓ 0.0264
average precision	0.9586	0.9647	↑ 0.0061
one-error	0.0070	0.0211	↑ 0.0141

("↓" means the metric value of IGAECM is lower than that of CARCTM, while "↑" represents the opposite case)

IV. CONCLUSION

IGAECM is proposed to carry out CVD oriented HDP&SP&PP categorization. IG is adopted to evaluate the relevance of HDP&SP&PP variables in detecting various CVDs. IGSS is designed to determine automatically the quantity and types of HDPs, SPs & PPs in each of groups as well as the number of groups in detecting disparate CVDs. A hierarchical probabilistic SVM classifier, site-measured CVD relevant data, and five multi-label evaluation metrics (hamming loss, coverage, ranking loss, average precision, and one-error) are used to verify the effectiveness and validness of such a disease-oriented categorization method. Compared with the previous CARTCM, IGAECM has better performance in hamming loss, coverage, ranking loss, and average precision. Even though IGAECM has a fair performance in one-error, it is still more intelligent and sound than CARTCM in constructing robust and efficient hierarchical classifier.

ACKNOWLEDGMENT

This work was supported in part by the Research Committee of University of Macau under Grant No. MYRG2014-00060-FST, and in part by the Science and Technology Development Fund (FDCT) of Macau S.A.R under Grant No. 016/2012/A1.

REFERENCES

[1] World Health Organization. [Online]. Available: http://www.who.int/gho/ncd/mortality_morbidity/cvd/en/
[2] J. Fu, X. M. Wang, M. S. Wang, and W. Liu, "Noninvasive acoustical analysis system of coronary heart disease," in *Proc. 16th Southern. Conf. Biomedical Engineering*, Biloxi, 1997, pp. 239-241.
[3] M. Giardina, A. Francisco, P. McCullagh, and R. Harper, "A supervised learning approach to predicting coronary heart disease complications in type 2 diabetes mellitus patients," presented at the *IEEE Symposium*, *BIBE*, Arlington, October16-18, 2006.
[4] Z. Jin, Y. Sun, and A. C. Cheng, "Predicting cardiovascular disease from real-time electrocardiographic monitoring: An adaptive machine learning approach on a cell phone," presented at the *Engineering in Medicine and Biology Society*, September 3-6, 2009.
[5] G. P. Liu, G. Z. Li, Y. L. Wang, and Y. Q. Wang, "Modelling of inquiry diagnosis for coronary heart disease in traditional Chinese medicine by using multi-label learning," presented at the *BMC Complementary and Alternative Medicine*, 2010.
[6] K. Minas, M. Joseph, and P. Constantinos, "Assessment of the risk of coronary heart event based on data mining," presented at *8th IEEE International Conference*, *BioInformatics and BioEngineering*, Athens, October 1-5, 2008.
[7] S. P. Ge, "Real-time 3D echocardiography for congenital heart disease: From fetus to adults," *PMPH-USA*, 2013.
[8] F. Javed, P. A. Venkatachalam, and A. F. M. Hani, "Knowledge based system with embedded intelligent heart sound analyser for diagnosing cardiovascular disorders," *Journal of Medical Engineering & Technology*, pp. 341-350, vol. 31, no. 5, 2007.
[9] F. F. Zhao, "Contemporary sphymology in traditional Chinese medicine," *People's Medical Publishing House*, Beijing China, pp. 227-240.
[10] M. You, G. Z. Li, "Medical diagnosis by using machine learning techniques," *Data Analytics for Traditional Chinese Medicine Research*, ch. 3, Springer, 2014, pp. 39-80.
[11] X. Liu, P. Lu, X. H. Zuo, Y. Gao, and J. X. Chen, "A new method of modeling of inquiry diagnosis for coronary heart disease in traditional Chinese medicine," presented at *4th International Conference*, *Biomedical Engineering and Informatics (BMEI)*, Shanghai, October 15-17, 2011.
[12] H. Shao, G. Z. Li, G. P. Liu, and Y. Q. Wang, "Symptom selection for multi-label data of inquiry diagnosis in traditional Chinese medicine," *Science China Information Sciences*, Springer, 2012, pp. 1-13.
[13] J. Shi, M. C. Dong, B. D. Sekar, and W. K. Lei, "Prognose coronary heart diseases through sphygmogram analysis and SVM classifier," presented at *Information, Communications and Signal Processing*, Macau, December 8-10, 2009.
[14] R. A. Johnson, "Miller & Freund's probability and statistics for engineers," in *Prentice-Hall Inc*., New Jersey, 2000.
[15] J. Shi, B. D. Sekar, M. C. Dong, and W. K. Lei, "Fuzzy neural networks to detect cardiovascular diseases hierarchically," presented at the*10th IEEE*, *Computer and Information Technology*, Bradford, pp. 703-708, 2010.
[16] M. B. Chen, T. C. Tang, J. L. Ma, and M. C. Dong, "CVD oriented HDP&PP categorization," *Future Information Technology*, Springer, Berlin, pp. 601-606, 2014.

[17] J. R. Quinlan, "Introduction of decision trees," *Machine Learning*, pp. 81-106, 1986.
[18] J. R. Quinlan, "C4.5: Programs for machine learning," *Machine Learning*, vol. 16, no. 3, pp. 235-240, 1993.
[19] L. Yu and H. Liu, "Feature selection for high-dimensional data: A fast correlation-based filter solution," in *Proc. the Twentieth International Conference on Machine Learning*, 2003, pp. 856-863.
[20] M. B. Chen, B. B. Fu, T. C. Tang, J. L. Ma, and M. C. Dong, "Hierarchical probabilistic support vector machine for detecting cardiovascular diseases," *International Journal of Bioscience, Biochemistry and Bioinformatics*, vol. 4, no. 5, pp. 340-344, 2014.
[21] J. S. Russell and P. Norvig, "Artificial intelligence: A modern approach," *Upper Saddle River*, Prentice Hall, Upper Saddle River, New Jersey, pp. 111-114, 2003.
[22] V. Vapnik, "The nature of statistical learning theory," in *Data Mining and Knowledge Discovery*, New York: Springer, 1995, pp. 1-47.
[23] C. C. Chang, C. J. Lin, "LIBSVM: A library for support vector machines," *Software*.
[24] R. E. Schapire and Y. Singer, "Boostexter: A boosting-based system for textcategorization," *Machine Learning*, vol. 39, no. 2-3, pp. 135-168, 2000.

MuBo Chen is currently pursuing MSc degree in ECE department of University of Macau, Macau. His current research is oriented to R&D a hybrid intelligent system to diagnose cardiovascular disease. He received Bachelor degree in CS in 2012 from Guangdong University of Foreign Studies. His research interests include machine learning, applying AI technology in various engineering and biomedical applications etc.

Ming Chui Dong is Full Professor and PhD supervisor of ECE Department of University of Macau, Automation Department of Tsinghua University, China and Professor of YanTai University, China. He received MSc degree in EEE in 1975 at Tsinghua University, China, Visiting Scholar in EEE in 1981 at Rome University, Italy. Hismain research interests are AI and its application in biomedical engineering, CIMS, fault diagnosis, AI in vision-to-text, voice-to-text machine translation etc. His postal address is Faculty of Science and Technology, University of Macau, Macau and his email address is mcdong@umac.mo, charley_dong@hotmail.com.