Hybrid discrimination method for samples classification in medicine

Xiaoyu Chen[1], Bo Liu[1,*], Xin Xia[1], Dandan Yan[1], Wang Yan[2,3], Lizhuang Ma[4]

[1.] *Department of Information Center*, East Hospital, Tongji University, School of Medicine, Jimo Road 150, Shanghai, China
[2.] *School of Continuing Education*, Shanghai Jiao Tong University, Shanghai, 200240, China
[3.] Provincial Key Laboratory for Computer Information Processing Technology, Soochow University, Suzhou, 215006, China
[4.] *Department of Computer Science*, Shanghai Jiao Tong University, Shanghai, 200240, China

***Abstract* — As indispensable solutions in classification problems, discrimination for samples has been employed in medicine, and it is performed subjectively by physicians at present, which hinders the diagnosis and treatment in medicine. In this paper, a hybrid discrimination method (HDM) in medicine is proposed, which consists of two phases, including attribute selection phase and discriminant phase. In attribute selection phase, critical attributes are selected from the original features by linear correlation and C5.0 decision tree. In discriminant phase, samples are discriminated by discriminant analysis. This discrimination method is evaluated through five datasets of chronic hepatitis B, cardiac Single Proton Emission Computed Tomography (SPECT) images, Lung Cancer, Hepatitis survival and Iris plant for demonstrating its viability and applications. Finally, this proposed method has obtained the critical clinical lab indicators and discriminants related to three syndromes in CHB dataset, and it also performs well than some typical classification methods in the other four datasets for its broader applications.**

Keywords - Hybrid discrimination, Attribute selection, Linear correlation, C5.0, Discriminant analysis

I. INTRODUCTION

With the rapid development of decade years, researches about applications of computer and information technology in medicine has become a new highlight direction, Medical information includes clinical medical information and hospital management information. Researchers are interested mainly in clinical medical information, as the objects of data mining researches, this part of medical data resources can reflect unique characters or rules of medical information. Modern medical datasets are composed of a great many attributes (symptoms and lab indicators), so it tends to use statistical models, data mining or other quantitative analysis methods in medical research [1-3], including irrelevant attributes and noise data. Attribute selection is frequently adopted to identify and remove the irrelevant and redundant information.

Fewer attributes means less data need to be collected, and collecting data is never an easy job because of time consuming, and the selection of appropriate subset of the available attributes can make a compact and easily interpretable representation of the target concept, it can improve the classification accuracy in medical region.

So attribute selection is an important step for data mining, and it is adopted to identify and remove the irrelevant and redundant information as much as possible. In recent years, some research efforts about syndrome discrimination have been acquired. For example, Qu et al. [4] used Decision tree method to self-extract diagnostic rules from 290 patients related to blood stasis syndrome. In result, 35 attributes from 52 features were selected to build classification model and five diagnostic rules were induced from the model; Zhang et al. [5] proposed hierarchical latent class (HLC) models and applied HLC models to discover the latent variables in the diagnosis of kidney deficiency syndromes, 2600 cases were investigated to collect 67 symptoms related to kidney deficiency syndromes, and the diagnosis based on the model made conclusions matches with those by experts; and there are also some technical applications of classification or decision-making methods in medical classification and decision [6-8].

In this paper, basing on the research datasets, this study manages to propose a hybrid discriminant method (HDM) for categories discrimination in classification of medical samples. It consists of two phases, attribute selection phase and discriminant phase. First, critical attributes pertaining to category labels of samples are selected through attribute selection. Second, samples are differentiated in discriminant phase. Five datasets including four medical datasets and one plant dataset are used for proving the viability and applications of HDM.

For showing its viability, the proposed HDM is evaluated on the chronic hepatitis B (CHB) dataset for main demonstration, and the CHB dataset includes 664 patients of three syndromes in traditional Chinese medicine and 83 clinical indicators, three syndromes are Damp Heat in the Liver and Gallbladder, Liver Qi Stagnation and Spleen Deficiency, and Yin Deficiency of Liver and Kidney, they are encoded with letters “A, B, C” for brief, with A represents Damp Heat in the Liver and Gallbladder, B for Liver Qi Stagnation and Spleen Deficiency, and C for Yin

* Address correspondence to the author at East Hospital, Tongji University,Shanghai, China;
Tel: +86 021 38804518-17280; Email: liubonew@126.com

Deficiency of Liver and Kidney. The nominal values of "A, B, C" are corresponding to "1, 2, 3" respectively.

In order to prove broader applications of HDM, it is also demonstrated on the datasets of cardiac Single Proton Emission Computed Tomography (SPECT) images, Lung cancer, Hepatitis survival and Iris plant, and these datasets originate from UCI datasets, which are downloaded at the website of http:// archive. ics.uci.edu/ml/datasets.html. The paper is organized as 4 parts: Section 2 describes the ideas of how to construct HDM. The experimental results based on HDM are shown in Section 3. Some discussions and comparisons about this proposed method are presented in Section 4. Section 5 gives the conclusion.

II. MATERIAL AND METHODOLOGY

A. Material

1) CHB Dataset

The CHB dataset originates from chronic hepatitis B patients, dating from November 2009 to April 2010. It contains 664 chronic hepatitis B cases and 83 clinical lab indicators as the items of information collection, including: routine tests of blood, urine, and liver function indicators, immune indicators, renal function indicators, blood glucose, and lipids, and the diagnostic criteria are referred to Prevention and treatment programs of Viral hepatitis [9] issued by the Chinese liver disease association and society of infectious diseases. The distributions of samples are partitioned into three types: 367 cases are syndrome A, 188 cases are syndrome B, and the other 109 cases are syndrome C.

The attributes are encoded by the following rules:

(i) In the dataset, some lab indicators are encoded with binary values (0, 1)；with 0 representing negative and 1 for positive; some indicators are encoded by the four-value ordinal scales measured by the level degree, with 0 representing normal level, 1 for slight level, 2 for medium level, and 3 for serious level; the others are encoded with numeric values.

(ii)The missing values of cases in this dataset are replaced by mean values of the corresponding attributes. Missing vales of all the attributes are less than 10%.

2) SPECT Dataset

The dataset originates from University of Colorado, it describes diagnosing of cardiac Single Proton Emission Computed Tomography (SPECT) images and has 267 instances that are described by 45 attributes. Each patient is classified into two categories: normal and abnormal. The database of 267 SPECT image sets (patients) is processed to extract attributes that summarize the original SPECT images. As a result, 44 continuous attribute is created for each patient, and the category attribute is a binary value (0, 1).

3) Lung Cancer Dataset

This dataset is used to illustrate the power of the optimal discriminant plane even in ill-posed settings. It describes 3 types of pathological lung cancers, includes 32 samples and 57 nominal attributes, the category attribute takes on integer values from 1 to 3.

4) Hepatitis Survival Dataset

This dataset was donated by Jozef Stefan Institute, it contains 155 samples and 20 attributes, including symptoms, signs, and lab indicators. Each patient is classified into two categories: live and die. Age and lab indicators are continuous values, and symptoms, signs, or category attribute are binary values (1, 2).

5) Iris Plant Dataset

This plant dataset was created by R.A. Fisher and donated by Michael Marshall, it may be the best known database to be found in the pattern recognition literature and it contains 5 attributes and 3 classes of 50 instances each, where each class refers to a type of iris plant. The category attribute is nominal; the other attributes are continuous values.

B. Methodology

Medical datasets are inevitable to contain plenty of information from redundant and irrelevant attributes, as to lower down efficacy and performance of data mining algorithms and cause incomprehensive results, that is called Hughes phenomenon [10]. Obviously, the appropriate subset of attributes can yield an accurate and interpretable result for focusing on the significant attributes objectively in syndrome differentiation. Therefore, attribute selection is a very important step for data mining and analyzing methods.

In HDM, we combine data mining and statistical methods together to construct a new method for classification; HDM consists of three classification methods of linear correlation, C5.0 decision tree and discriminant analysis as its components. Among them, linear correlation and discriminant analysis are linear classification methods, and C5.0 decision tree is a non-linear classification method. Linear correlation with C5.0 decision tree together to select critical attributes as the first phase, and in the second phase, each case is discriminated basing on selected critical attributes by discriminant analysis. And we take CHB dataset for instance; the logic process of HDM for instructing its construction is shown in Fig. 1.

1) Linear Correlation

Linear correlation is a method based on symmetric correlation measurement, firstly, it can remove those attributes which are in low relationship with the decision attribute (i.e., the linear correlation coefficient between condition attributes and the decision attribute is close to 0.), and secondly, after the removal of some condition attributes

in high correlations with the decision attribute, the redundancy between attributes can be reduced and it will not damage the prediction by characteristic attributes to the decision attribute [11]. For a random variable (X, Y), the linear correlation coefficient between them can be obtained by Eq. (1) as follow:

$$r = \frac{\sum_i (x_i - \overline{x_i})(y_i - \overline{y_i})}{\sqrt{\sum_i (x_i - \overline{x_i})^2}\sqrt{\sum_i (y_i - \overline{y_i})^2}} \quad (1)$$

In which is the mean value of variable X, is the mean value of variable Y, and the value of correlation coefficient r is between [-1,1]. If the variables X and Y are in complete correlation, r is 1or – 1, if the variables X and Y are completely independent, r is equal to 0. Basing on the correlation coefficient r of linear correlation, we use the concept of "importance index" (i.e. importance index=1-P, P is the probability of complete irrelevance of this independent variable with the target variable, its value is corresponding to correlation coefficient r in statistical methods, such as Chi-square test or F- test.) to calculate the importance of each condition attribute to the decision attribute, then the most important condition attributes (importance index >0.9) will be in reservation and others are filtered out [12].

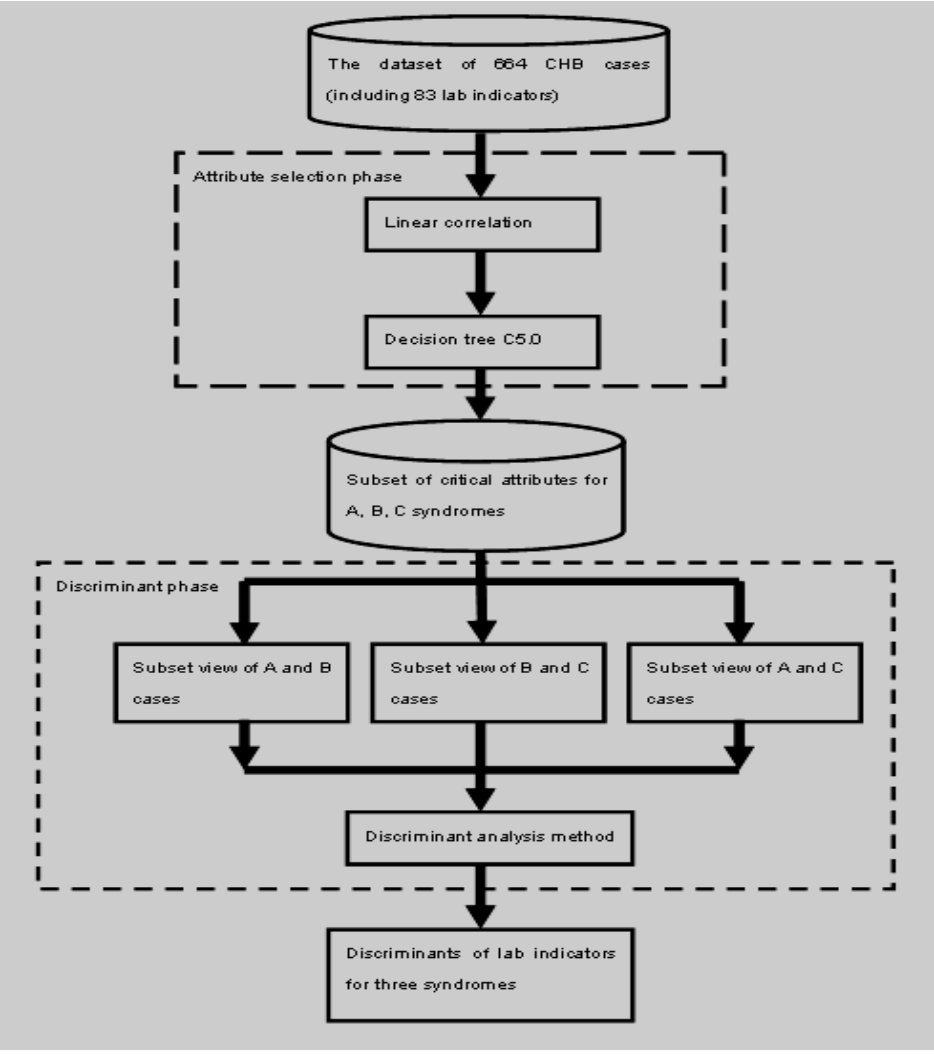

Fig.1 Logical Process for Syndrome Discrimination

2) C5.0 Tree

As an indispensable technology of data mining, decision tree has been applied widely in classification, prediction, rules extraction and other areas to solve the key issues of data classification. The fundamental idea of decision tree is to find the decision attribute through top-down recursive way and depending on the proper value to determine the node down from the branch and acquire the conclusion in

leaf nodes of the tree. That is partitioning the training set recursively, until all the records of each subset belong to one class, or the predominant majority of each subset belong to one class. So each path from the root to a leaf node corresponds to a conjunctive rule, and the whole decision tree corresponds to a group of extraction expression rules. (Relevant algorithm is shown in Fig. 2). In Fig. 2, we can see that partitioning method Fn is the key of decision tree algorithm, different Fn methods represent different decision tree algorithms, which are more representative and in broader applications among them, such as ID3 [13], C4.5 [14] based on information entropy. In this study, as the commercial version of C4.5, C5.0 decision tree improves the aspects of generating rules and algorithm precision to achieve more accurate generation rules, faster speed and lower error rate, and it is more suitable for classification of large data sets [12].

```
Input node n, Dataset D, partitioning method of Fn ();
Output Decision tree on node n as its root node, Dataset
D and partitioning method of Fn ();
1  Procedure Build Tree
2   Initialization root node n
3   Find decision feature n according to Fn on Dataset
D
4   If node n meets with conditions of partition
5   According to decision feature, Partition dataset D
    into D1 and D2, and creates two sub nodes of n,
    namely n1 and n2;
6    Build Tree (n1, D1, Fn);
7    Build Tree (n2, D2, Fn);
8   End if
```

Fig. 2 Formation of Decision Tree

3) Discriminant Analysis

Discriminant analysis [15] is based on an available number of samples classified in a number of clear indicators of the observation, and set up a discriminant function and rule on the indicators for classification. Then classify new samples into two types A and B according to the discriminant function and rule to make the lowest mistake classification rate. Methods of discriminant analysis can be divided into: fisher, maximum likelihood, Bayes formula and gradual selection or full model discriminant analysis [15]. Basing on above methods, in this paper, we use full model of fisher discriminant analysis, and it is shown by Eq. (2).

$$Z = C_1 X_1 + C_2 X_2 + \ldots + C_m X_m \tag{2}$$

In this equation, Ci is the coefficients of the function, and Xi is the indicators, samples are divided into two groups through calculating the values of this function, if values of Z are bigger than the cutoff value ZC for classification, samples are divided into type A, if values of Z are smaller than ZC, samples are classified as type B, and if values of Z are equal to ZC, samples cannot be classified between A and B. And the equation of ZC is listed in Eq. (3) below:

$$ZC = \frac{\overline{Z}_A + \overline{Z}_B}{2} \tag{3}$$

III. RESULTS

A. HDM on CHB Dataset

HDM includes two phases, attribute selection phase and discriminant phase. Attribute selection phase of HDM is based on 83 clinical indicators, and it consists of two parts, including linear correlation and C5.0 decision tree, during this phase, irrelevant attributes to the syndrome category are filtered out through two parts above. In discriminant phase, samples are discriminated into different categories through discriminant analysis.

1) Linear Correlation

To calculate the value of importance index of each attribute related to the category attribute on the dataset of 664 cases, we find 16 attributes of importance index greater than 0.9 (i.e. P<0.1.) from 83 clinical indicators. The 16 attributes are listed in Table I.

2) C5.0 Tree

Basing on attributes acquired from the linear correlation, decision tree of the 16 lab indicators is constructed by C5.0 decision tree method. In the 16 importance indexes of lab indicators decision tree, 10 attributes are selected as follows: Pre S1 antibody, Pre S1 antigen, LDL-C, APTT, basophil, CD4+, uric acid, ALT, percentage of basophil, PT. Among them, Pre S1 protein antibody and Pre S1 antigen are discrete attributes, others are continuous attributes, and the decision tree is shown in Fig. 3, and it has obtained a total classification accuracy of 69.13%.

3) Discriminant analysis

The subset of 10 attributes (mentioned in 3.1.2) is selected by C5.0 decision tree. In this phase, considering that most attributes obtained from decision tree are numeric values, so we apply full model of fisher discriminant analysis method to establish discriminants of syndrome discrimination by these attributes. Finally, we can determine the correlation coefficients of each attribute selected and the cutoff values for syndrome discrimination between A-B, B-C and A-C in three syndromes. Basing on 10 critical attributes, classification accuracies and cutoff values of the discriminants ZA, B(for A-B), ZB, C(for B-C), and ZA, C(for A-C) are listed in Table II, the coefficients of each attribute in discriminants and expressions of these discriminants for three syndromes are shown in Table III and Fig. 4 respectively.

TABLE I.THE IMPORTANT ATTRIBUTES SELECTED FROM LINEAR CORRELATION

Attribute name	importance	Attribute name	importance
Pre S1 antibody	1.0	LDL-C	0.985
percentage of basophil	1.0	ALP	0.984
PT	1.0	percentage of monocyte	0.981
APTT	1.0	ALT	0.978
Pre S1 antigen	0.999	Cr	0.972
uric acid	0.998	CD4+	0.928
TC	0.991	basophil	0.92
HDL-C	0.986	hemoglobin	0.904

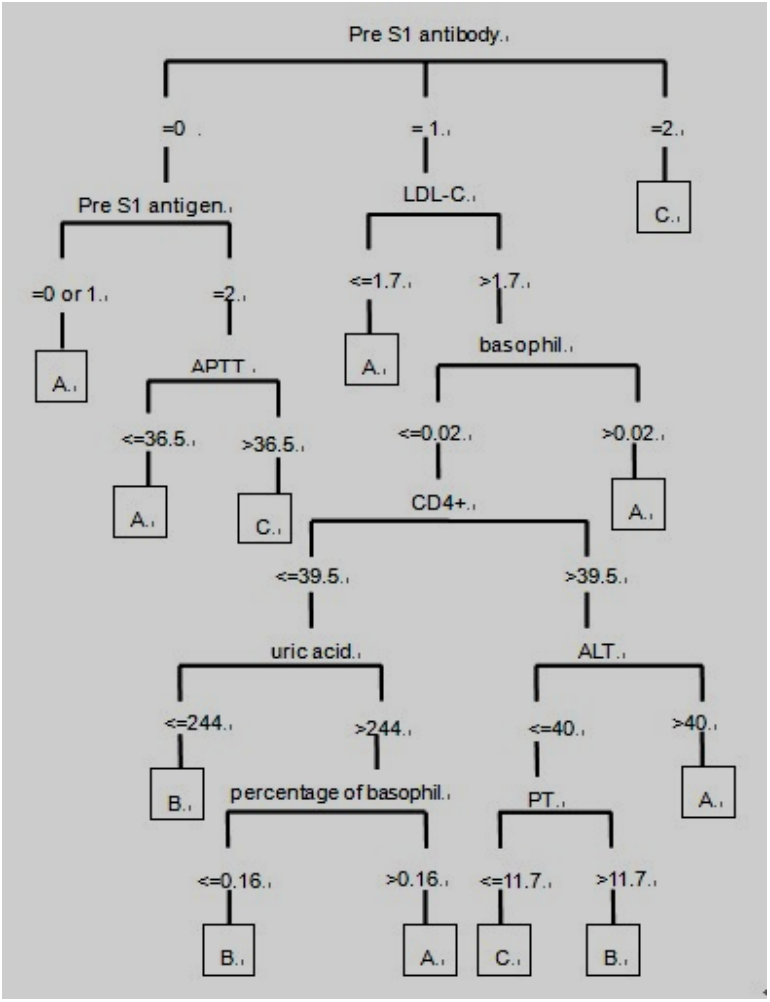

Fig.3 Decision Tree Based on Lab Indicators

TABLE II. CLASSIFICATION ACCURACIES AND CUTOFF VALUES OF DISCRIMINANTS

Discriminant	Classification Accuracy	Cutoff values
$Z_{A, B}$	61.98%	-0.0895
$Z_{B, C}$	65.20%	-0.111
$Z_{A, C}$	66.67%	-0.1995

TABLE III. COEFFICIENTS OF ATTRIBUTES IN DISCRIMINANTS

Attribute name	$Z_{A,B}$	$Z_{B,C}$	$Z_{A,C}$
ALT	-0.08	0.194	0.235
PT	-0.167	0.411	0.307
APTT	-0.481	-0.085	-0.387
Pre S1 antigen	0.141	0.2	0.256
Pre S1 antibody	-0.241	0.451	0.363
CD4	0.283	-0.348	-0.206
uric acid	0.544	-0.372	-0.058
LDL-C	-0.098	-0.185	-0.249
basophil	0.079	0.089	0.079
basophil percentage	0.567	-0.209	0.098

$Z_{A,B}$ = -0.08 ALT- 0.167 PT- 0.481 APTT+ 0.141 Pre S1 antigen- 0.241 Pre S1 antibody+ 0.283 CD4+ 0.544 uric acid- 0.098 LDL-C+ 0.079 basophil+ 0.567 basophil percentage	$Z_{B,C}$ = 0.194 ALT+ 0.411 PT- 0.085 APTT+ 0.2 Pre S1 antigen+ 0.451 Pre S1 antibody- 0.348 CD4- 0.372 uric acid- 0.185 LDL-C+ 0.089 basophil- 0.209 basophil percentage	$Z_{A,C}$ = 0.235 ALT+ 0.307 PT- 0.387 APTT+ 0.256 Pre S1 antigen+ 0.363 Pre S1 antibody- 0.206 CD4- 0.058 uric acid- 0.249 LDL-C+ 0.079 basophil+ 0.098 basophil percentage
If $Z_{A,B}$ > -0.0895, cases are identifies as A syndrome. else if $Z_{A,B}$ < -0.0895, cases are identified as B syndrome else no syndrome can be indentified between A and B.	If $Z_{B,C}$ > -0.111, cases are identifies as B syndrome. else if $Z_{B,C}$ < -0.111, cases are identified as C syndrome else no syndrome can be indentified between B and C	If $Z_{A,C}$ > -0.1995, cases are identifies as A syndrome. else if $Z_{A,C}$ < -0.1995, cases are identified as C syndrome else no syndrome can be indentified between A and C.

Fig. 4 Discriminant Expressions for A, B and C Syndrome

B. HDM on UCI Datasets

On four UCI datasets of SPECT images, Lung Cancer, Hepatitis survival, and Iris plant, HDM can also perform well in samples classification. The classification accuracies of HDM on the UCI datasets are shown in Table IV, and we can see that HDM can also perform well not only in classification of medical datasets.

C. Comparison with Typical Classification Methods

1) Comparison on CHB Dataset

In this experiment, as a proposed classifier, HDM is compared with six typical classifiers on CHB dataset, the classification accuracies are shown in Table V and Fig. 5. And these experimental results show that although the classification accuracies are not very high, they are better than those typical classifiers.

2) Comparison on UCI Datasets

The differentiation of analysis approach plays an important role in the data classification or infusion [16-18]. For further verification of better efficacy in classification and broader applications, the performance of HDM is compared with six typical methods above respectively using these four UCI datasets. Classification accuracies of HDM and the methods above are shown in Fig. 6, it shows that the classification accuracy with HDM outperforms well than those of above methods in total. Therefore, the proposed HDM method can be effectively applied into other different fields.

TABLE IV. CLASSIFICATION ACCURACY(%)OF HDM ON UCI DATASETS

	SPECTF	Lung Cancer	Hepatitis survival	Iris plant
HDM	76.3	75.0	84.4	96.0

TABLE V. CLASSIFICATION ACCURACY(%) COMPARISON ON CHB DATASET

Methods	A-B	B-C	A-C	Total classification
HTCMDM	61.98	65.20	66.67	65.51
Naïve Bayes	17.65	47.81	21.84	25.90
Complement Naïve Bayes	22.34	38.38	29.41	28.46
BayesNet	61.80	13.47	63.65	51.57
Logistic	56.04	25.58	57.73	49.84
LogitBoost	55.31	27.95	63.65	56.02
REPTree	59.09	13.80	64.49	50.90

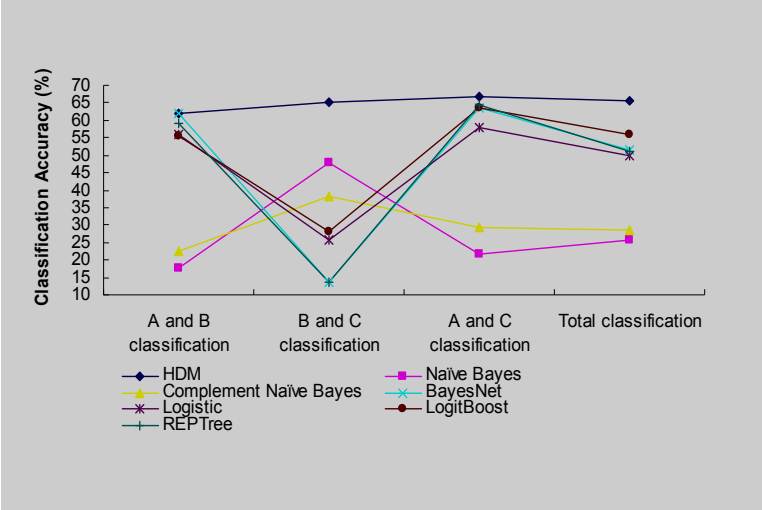

Fig.5 Comparison Between HDM and Other Methods on CHB Dataset

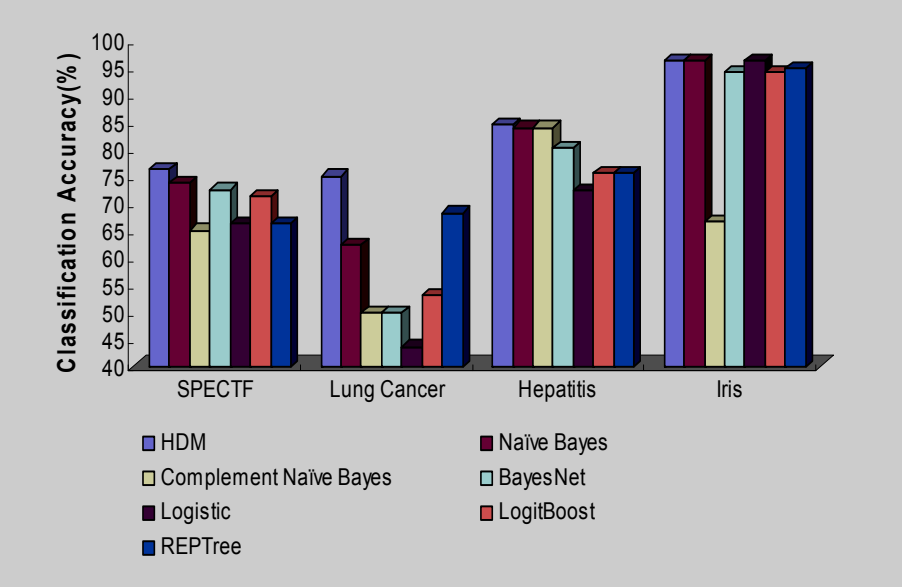

Fig. 6 Comparison Between HDM and Above Methods on UCI Datasets

IV. DISCUSSION

The proposed hybrid discriminant method (HDM) for samples classification is demonstrated on four medical datasets and one iris plant dataset for viability and applications. From the results of classification in Section 3, we can conclude that the proposed HDM can be applied in samples discrimination in medicine and other fields.

On datasets of cardiac Single Proton Emission Computed Tomography (SPECT) images, Lung Cancer, Hepatitis survival and Iris plant from UCI datasets, HDM has acquired an accuracy of 76%, 75%, 84.4%, and 96% respectively in samples classification. On CHB dataset, 10 lab indicators (in Table I.) are selected through linear correlation and decision tree, including Pre S1 antibody, Pre S1 antigen, LDL-C, APTT, basophil, CD4+, uric acid, ALT, percentage of basophil and PT. And we have also acquired discriminant expressions of clinical syndrome discrimination of CHB, and from the three cutoff values and discriminant expressions in Fig. 4, we can summarize the relationship between the cutoff values of these expressions and three syndrome types. It is expressed by the variable Z Syndrome and listed in Table VI. below. Through discriminant expressions and correlated coefficients (in Table III and Fig. 4) for syndrome differentiation between syndromes of A, B, and C by full model discriminant analysis, in spite of classification accuracies between A-B, B-C and A-C syndromes listed in Table II are not very high, they are in equilibrium relatively. It shows that although the full model of discriminant analysis is fit for conducting a comprehensive analysis of variables selected in the study, this method determines all attributes selected as the independent variables of the discriminant function without considering each variable's significance or contribution to the discriminant function result, and it will cause a relative bigger bias and may have an impact on the classification accuracy of syndrome differentiation in some extent. This dataset is based on lab indicators of 664 cases, and lab indicators are different from symptoms or signs in syndromes, different syndromes contain distinct symptoms or signs, but all clinical lab indicators are shared by various syndromes in CHB. In spite of lab indicators are not so specific as symptoms, they are more objectively observed and widely used by physicians in clinics.

Compared with six typical classifiers of Naïve Bayes, Complement Naïve Bayes, BayesNet, Logistic, LogitBoost, and REPTree, HDM has performed better than these typical methods in classification on four medical datasets and one iris plant dataset for its broader applications.

TABLE VI. THE RELATIONSHIP BETWEEN EXPRESSION VALUES AND SYNDROME TYPE

$Z_{Syndrome}$	Syndrome type
$Z_{Syndrome} > -0.0895$	A
$-0.111 < Z_{Syndrome} < -0.0895$	B
$-0.1995 < Z_{Syndrome} < -0.111$	A or C
$Z_{Syndrome} < -0.1995$	C

V. CONCLUSION

In this paper, a hybrid discriminant method (HDM) is proposed for samples classification as a new method. It combines linear (linear correlation and discriminant analysis) and non-linear (C5.0 decision tree) classification methods together, and its viability and applications are demonstrated on four medical datasets and one iris plant dataset. Through the experimental classification results, we can conclude that HDM outperforms six typical classifiers, and it can not only be applied in medicine, but also in other fields. And in future, the proposed HDM will be improved for higher classification accuracies and more optimal attribute subsets.

CONFLICT OF INTEREST

The authors confirm that this article content has no conflicts of interest.

ACKNOWLEDGMENT

The research is supported Open Research Fund of Jiangsu Provincial Key Laboratory for Computer Information Processing Technology (No.KJS1226), it is also acknowledged for the support from the Natural Science Foundation of Shanghai City (No.14ZR1422700), and the data of the research for analysis partly originates from National Science and Technology Major Project of China (2009ZX10004-601).

REFERENCES

[1] W. Lu, "Relationship of traditional Chinese medicine syndromes and virus nucleic acid indicators in chronic hepatitis B", Hubei Journal of Traditional Chinese Medicine. vol.28 ,no.7, pp.16, July 2006.

[2] B. He, S. Mao, "Relations between antiviral efficacy of lamivudine and traditional Chinese medicine syndromes in chronic hepatitis B", Zhejiang Journal of Medicine. vol.14 ,no.1, pp. 15-16 ,January 2004.

[3] W. Xiao, S. Ju, J. Zhu et al., "A data-mining algorithm oriented to the largest two-dimensional frequent item set in the field of traditional Chinese medicine" , Journal of Chinese Computer Systems. vol.28, no.12, pp. 2193-2198, December 2007. .

[4] H.Qu, L.Mao, 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. 709–711, 727, June 2005.

[5] N.L.Zhang, S.Yuan, T.Chen, Y.Wang, "Latent tree models and diagnosis in traditional Chinese medicine", Artificial Intelligence in Medicine, vol. 42, pp. 229–245, March 2008.

[6] Casey C. Bennett, Kris Hauser, "Artificial intelligence framework for simulating clinical decision-making: A Markov decision process approach", Artificial Intelligence in Medicine, vol.57, pp.9-19, December 2012.

[7] Thomas G. Kannampallil, Amy Franklin, Rashmi Mishra, Khalid F. Almoosa, Trevor Cohen, Vimla L. Patel, "Understanding the nature of information seeking behavior in critical care: Implications for the design of health information technology", Artificial Intelligence in Medicine, vol.57, pp.21-29, December 2012.

[8] Hwan-Jeu Yu, Hong-Shiee Lai, Kuo-Hsin Chen et al." A sharable cloud-based pancreaticoduodenectomy collaborative database for physicians: Emphasis on security and clinical rule supporting", computer methods and programs in biomedicine, vol.3, pp. 488-497, April 2013.

[9] Chinese Medical Association, "Prevention and treatment programs of viral hepatitis", Chinese Journal of Internal Medicine. vol.40, no.1, pp. 62-68, January 2001.

[10] G.F.Hughes, "On the mean accuracy of statistical pattern recognizers", Information Theory. vol.14, pp. 55-63, January 1968.

[11] S.K.Das," Feature selection with a linear dependence measure", IEEE Transactions on Computers. vol. 20, pp. 1106- 1109, September 1971.

[12] P.Xiong, Data mining algorithms and practice of Clementine, Tsinghua University Press, Beijing, 2011.

[13] J.R.Quinlan, "Induction of decision trees", Machine Learning. vol.1, pp. 81–106, January 1986.

[14] J.R.Quinlan, C4.5: Programs for machine learning. Morgan Kaufmann Publishers, San Francisco, 1993.

[15] J.L.Xia, H.W.Jiang, Discriminant analysis and clustering analysis, in: H.Yan, Y.Y. Xu, Medical statistics, People's Medical Publishing House, Beijing, 2005, pp. 373 -374.

[16] J. H. Wang and K. Liu, "Feature-based fuzzy neural network approach for intrusion data classification", Journal of Computational Intelligence and Electronic Systems, vol.1, pp.99-103, June 2012.

[17] S. A. Quadri and O. Sidek," Role of algorithm engineering in data fusion algorithms", Journal of Computational Intelligence and Electronic Systems, vol.2, pp.29-35, June 2013.

[18] M. Göndör, V. P. Bresfelean, "REPTree and M5P for Measuring Fiscal Policy Influences on the Romanian Capital Market during 2003-2010", International Journal of Mathematics and Computers in Stimulation, vol. 6, pp. 378–386, June 2012.