Research on CVDs Prediction and Early Warning Techniques in Healthcare Monitoring System

Yi Chai[1], Guixia Kang[1], Ningbo Zhang[1], Jianwei Wu[1], Xiaoshuang Liu[1], Yuncheng Liu[2]
[1] Key Laboratory of Universal Wireless Communications, Ministry of Education
Beijing University of Posts and Telecommunications, Beijing, China
Emails:zzu09cy@126.com,gxkang@bupt.edu.cn
[2] TGLD Information Center, Beijing China

***Abstract*—Chronic diseases are gradually becoming the principal factors of harm to people's health. Fortunately, the development of e-health provides a novel thought for chronic disease prevention and treatment. This paper focuses on the research of cardiovascular disease (CVDs) prevention and early warning techniques using e-health and data mining. In this paper, we will use weighted associative classification algorithm to model the data in healthcare database to determine the level of cardiovascular risk. Besides, on the basis of data mining and knowledge discovery, intelligent warning mechanisms are proposed to provide different services to patients with different levels of risk. The experimental results show that the used classification algorithm is a more effective mining algorithm in the field of healthcare with higher accuracy and better comprehension. Our study is of definite significance to help control risk level of CVDs patients.**

***Keywords*—Classification, CVDs, Fuzzy logic, Associative classifier**

I. Introduction

Cardiovascular diseases (CVDs) are currently the number one cause of death globally. According to statistics, in 2009 China's total population CVDs mortality rate is 236.8 per 100 thousand, accounting for 40.6% of the total death [1]. A large number of epidemiological and clinical studies have demonstrated that high morbidity and high mortality of CVDs have a close relationship with CVDs risk factors of the epidemic. As the pathogenesis of chronic diseases are complex and variable, it is yet difficult to make an accurate diagnosis in advance. However, the occurring and progressing of chronic diseases follows certain rules, and the harm can be significantly reduced by evaluating the patient's condition and taking pertinent intervention.

In recent years, wireless e-health (We-health) technology [2] developments provide a method for daily monitoring and efficient prevention of chronic disease. Similar to other healthcare monitoring system [3-4], our CVDs monitoring system shown in Fig.1, is built using We-health technology and ICT. CVDs related medical data are gathered by portable measuring devices in home environment and then transmitted via the home gateways through the Internet or 2G/3G network to the storage center and processing center. Building on the arriving data, as well as the related contexts (such as Medical history, family history and etc.) which have already been acquired and stored in the storage center, a predicting unit and a decision making unit will function to give out their respective results. Medical staffs can also timely check the results through a service platform and give health guidance through a reverse link.

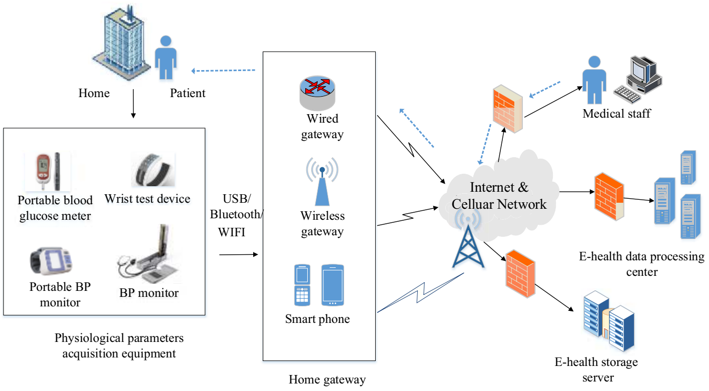

Fig.1. Architecture of CVDs monitoring system

In the process of long term application of CVDs monitoring system, massive practical chronic disease data containing potential knowledge have been accumulated. Doctors are very eager to extract valuable information from medical data, to discover the yet unknown rules, to optimize therapies and improve medical services. Practice has proven that useful conclusion can be effectively drawn using data mining (DM) techniques. Therefore, in perspective of preventive medicine, it is necessary to build predictive models with DM techniques to help doctors with diagnosis and clinical supervision. Data mining is a technology which is put forward under the background of "information explosion, knowledge poor" .Until now, classical data mining algorithms such as decision tree, neural network have been widely applied to many fields, including medical field.

This work was supported by Important National Science & Technology Specific Projects: No.2013ZX03006001, International Science and Technology Cooperation Project: No.2010DFA11590, and New Century Excellent Talents in University (NCET): NCET-11-0593.

978-1-4799-6644-8/14/$31.00 ©2014 IEEE

In this paper, we mainly research on the core techniques in the prediction unit and decision making unit of CVDs monitoring system. As for the prediction unit, we use the classification algorithm to model the actual data to determine the level of CVDs risk. In terms of understandability and accuracy, Association Classification is selected instead of the traditional classification algorithms. In the decision making unit, intelligent warning mechanisms are proposed to provide different services to patients with different levels of risk on the basis of data mining and knowledge discovery. Besides, in order to improve the accuracy of intelligent warning mechanism, fuzzy logic is also introduced to solve the problem of "sharp boundary" which is caused by the continuous variable in the process of decision making.

The whole paper is organized as follows: in Section II, the DM schemes and necessary input and output parameters are designed on the basis of authoritative medical guidelines of hypertension and CVDs [5-6]; in Section III, the Weighted Association Classification algorithm is discussed in detail, based on actual medical data; in Section IV, a novel early warning mechanism using fuzzy logic is put forward; and in Section V, experiments are conducted and results are analyzed ; The whole paper is concluded in the Section VI.

II. Design of DM Scheme for CVDs

A. Classification of Blood Pressure (BP)

According to authoritative medical guidelines for CVDs [5], hypertension is the most important risk factor for stroke onset among the Chinese people. So, it is necessary to pay attention to blood pressure levels of cardiovascular patients. Systolic Blood Pressure (SBP) and Diastolic Blood Pressure (DBP) are the two important parameters to measure BP level. The categorization of BP level, determined by the 2010 Hypertension Prevention and Treatment Guide of China [6], is shown in TABLE I. Particularly, if a patient's SBP and DBP are in different levels, then the final result will be the higher one.

TABLE I. Classification of BP level [6]

Classification	SBP (mmHg)	DBP (mmHg)
Normal	< 120	< 80
High-normal	120 ~ 139	80 ~ 89
Level 1 Hypertension	140 ~ 159	90 ~ 99
Level 2 Hypertension	160 ~ 179	100 ~ 109
Level 3 Hypertension	≥ 180	≥ 110
Isolated Systolic Hypertension(ISH)	≥ 140	< 90

B. Classification of Cardiovascular Risk

There are many common risk factors for cardiovascular disease except blood pleasure. The main risk factors in the prognosis of CVDs are shown in TABLE II. Besides a high BP level, target organ damage and other related diseases are also the signs of CVDs.

TABLE II. Main risk factor in prognosis of CVDs[5]

CCVD Risk Factor	Target Organ Damage	Other complication
Hypertension (Level 1~3) Male over 55, female over 65 Smoking history Impaired Glucose Tolerance Dyslipidemia Family antecedents of premature CCVDs Abdominal obesity (BMI ≥28kg/m^2) …	Left ventricular hypertrophy Atherosclerotic plaque eGFR $<$ 60 ml·min^{-1}·1.73m^{-2} Microalbuminuria …	Cerebrovascular disease Cardiac disease Renal disease Peripheral vascular disease Retinal disease Diabetes …

CVDs are the result of multiple risk factors combined action, therefore the risk of CVDs depends on not only the severity of a certain risk factors, but also the number and degree of risk factors at the same time. According to CVDs guideline, there are 4 cardiovascular risk levels, i.e. low, moderate, high and ultrahigh. But the clinical therapy for high risk and ultrahigh risk are both instant drug using. To simplify classification, these two levels can be combined as the new "high". The simplified classification of cardiovascular risk is given in TABLE III. In CVDs management, patients with moderate risk should be in close observation of BP, risk factors and target organ damage to decide when to use drug therapy. For patients at low risk, long-term observation, frequent BP checking and necessary intervention are needed.

TABLE III. Simplified classification of CVDs risk [6]

Condition	BP Level		
	Level 1	Level 2	Level 3
No CVDs risk factors	Low risk	Moderate risk	High risk
1~2 CVDs risk factors	Moderate risk	Moderate risk	High risk
3 CVDs risk factors, or target organ damage, or complications	High risk	High risk	High risk

C. DM Workflow and Input/ Output Parameters

Based on a patient's BP level and other conditions, we can give his or her classification result of CVDs risk. According to the general process for data mining problems, workflow for CVDs risk level classification is designed, shown in Fig. 2. The first step is data preparation. The original data are those collected by medical sensors and the prognosis data include the condition of CVDs risk factors, target organ damage and other complications. The next is data pre-processing, including screening, cleaning, converting, filling the missing and etc.

Generally, it usually takes most of the time. After data pre-processing, we will use an appropriate and high performance data mining algorithm to model the CVDs data, and finally get a general model which can be applied to new data. When designing the system, we refer to the Guidelines for Chinese Guide lines for Cardiovascular Disease Prevention [5], but the process can also be transplanted to other countries.

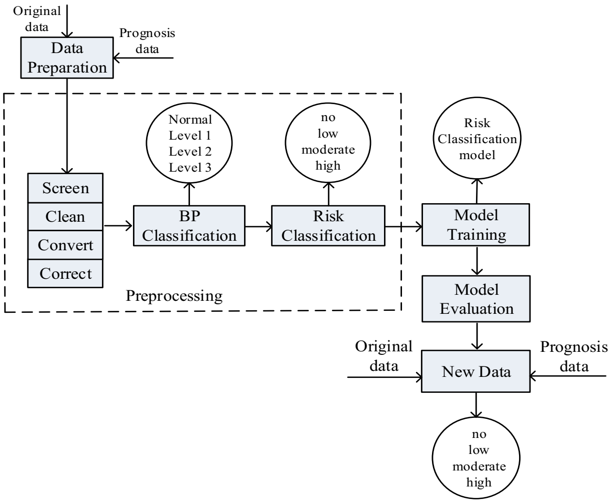

Fig.2. DM workflow of CVDs.

The input and output parameters of DM are shown in TABEL IV. There are two kinds of input data, numeric and categorical type. The output parameter is categorical type, which gives the patient's cardiovascular risk classification.

TABLE IV. DM PARAMETERS OF CVDS [5]

Type	Input	Output
NUMERIC	SBP, DBP, age, height, weight, eGFR	/
CATEGORICAL	gender, smoking history, impaired glucose tolerance, dyslipidemia,family antecedents of premature CVDs,left ventricular hypertrophy,atherosclerotic plaque, microalbuminuria, cerebrovascular disease, cardiac disease , etc	classification of cardiovascular risk

III. CVDS RISK LEVEL PREDICTION USING WEIGHTED ASSOCIATIVE CLASSIFIER

A. Methodology

Associative classification (AC) [7] is a rewarding technique that applies the methodology of association rule mining into classification and achieves higher classification accuracy and better robustness than traditional classification algorithms. Moreover, many rules found by AC cannot be discovered by traditional classification algorithms. Similar to other classifications, the whole process of AC consists of three steps in general: data preparation, model training and new data prediction. Specially, Associative classification training phase contains two major steps: (1) mining the specified minimum support threshold and minimum confidence threshold of associative classification rules; (2) clip redundancy, noise rules and construct classifiers.

However, Weighted Associative Classifier (WAC) is another concept proposed by Sunita Soni et al. [9]. It differs from traditional associative classifiers and assigns different weights to different attributes to get more accuracy in predictive system of some expertise fields such as medical field and etc. The whole process of WAC is shown in Fig.3 and the important details will be described as follows.

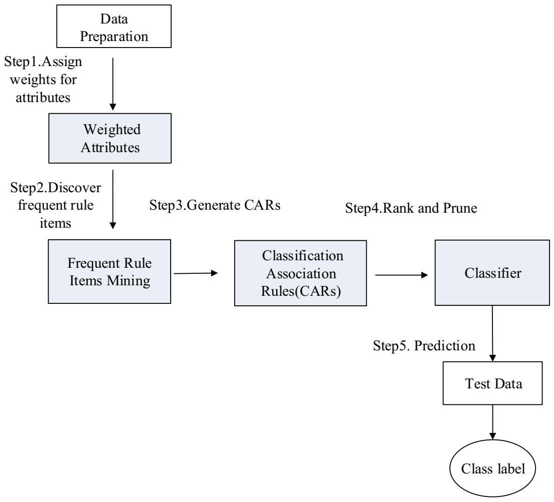
Data Preparation
Step1.Assign weights for attributes
Weighted Attributes
Step2.Discover frequent rule items
Step3.Generate CARs
Step4.Rank and Prune
Frequent Rule Items Mining
Classification Association Rules(CARs)
Classifier
Step5. Prediction
Test Data
Class label

Fig.3. Classification process of WAC [8]

First of all, it is essential to preprocess the data in the medical database to make it adapt to the mining process. Because associative classification is only suitable for processing categorical attributes, but medical data contain a lot of continuous attributes. Therefore, a series of discretization should be conducted at first.

Then, assign corresponding weight to each attribute. Each attribute is specified a weight ranging from 0 to 1 to represent the predicting capability of the attribute to the class label. Attributes that have more impact on class label will be assigned a higher weight, and attributes having less impact will be assigned a lower weight.

The following is to use Weighted Association Rule Mining Algorithm to generate an interesting pattern. In this phase, redefined weighted support and confidence are used rather than the traditional support and confidence [9].After processing, CARs (known as Classification Association Rules) are generated and are represented like $X \rightarrow Y$, where X is set of attributes and Y is the CVDs risk label. For an instance, {(Age, ">52"), (Smoking, "yes"), (Hypertension, "yes")} → CVDs risk ="high risk".

Lastly, store the rules in the rule base. The CARs from the rule base can be utilized to predict the class label, whenever a new patient's record is provided.

B. *Detailed Discussion of Classifier Combined with the CVDs Data*

Although Jyoti Soni et al. developed an intelligent and effective heart disease prediction system using WAC, they employed the data from the UCI machine learning dataset instead of the actual data collected from the chronic disease monitoring system [10].So we use the WAC to mine our collected CVDs data for interesting patterns and to verify better accuracy of WAC all at once.

1) Assign Attribute Weight

Attribute weight is assigned depending upon the domain. In the medical domain, symptoms can be allocated weights on the advice of the medical experts. As shown in Table V, all of the attributes in CVDs database are given weights according to the contribution degree to CVDs, which is measured by experienced Chinese experts in the field of CVDs. Maybe, they are not applicable to anyone in anyplace. So, adjusting measures to local conditions is sometimes necessary.

TABLE V. WEIGHT OF SYMPTOMS FOR CVDS

No.	Attributes	Symptoms	Weights
1	Age	<=40	0.1
2		>40&<55	0.3
3		>=55	0.5
4	Smoking habits	yes	0.7
5		no	0.6
6	Hypertension	yes	0.8
7		no	0.7
8	BMI	<=25	0.1
9		>=26&<=30	0.3
10		>=31	0.6
11	Dyslipidemia	yes	0.8
12		no	0.7
13	Diabetes	yes	0.7
14		no	0.6

2) Calculate Tuple /Record Weight

In view of data in a relational table, tuple weight can be calculated according to each attribute weight. It is average weight of all the attributes in the tuple. Considering that there are N attributes in the table, then the tuple weight can be denoted by $W_t = \sum_{i=1}^{N} weight(a_i)/N$, that is weighted average of N attribute weights [9]. Concrete examples are given to strengthen understanding in Table VI.

TABLE VI. TUPLE WEIGHT FOR CVDS

No.	Age	Smoking	Hypertension	BMI	Dyslipidemia	Diabetes
1	0.1	0.7	0.7	0.3	0.7	0.6
	Tuple weight		(0.1+0.7+0.7+0.3+0.7+0.6)/6=0.52			
2	0.3	0.7	0.8	0.1	0.7	0.6
	Tuple weight		(0.3+0.7+0.8+0.1+0.7+0.6)/6=0.53			

3) Weighted Support

Weighted support (WSP) of a rule $X \rightarrow Y$, where Y denotes the class label and X denotes frequent value item set, is the fraction of weight of the tuples that contain frequent item value set X and class label Y relative to the weight of all transactions [9]. For example, considering that a rule R (Diabetes ="yes" → CVDs_label ="high") then weighted support of R is calculated as:

$$WSP(R) = \frac{\begin{array}{c}\textit{sum of Tuple Weight having the condition}\\ \textit{Diabetes="yes" and also given CVDs\_label = "high\_risk"}\end{array}}{\textit{sum of weight of all transactions}} \quad (1).$$

4) Weighted Confidence

Based on the definition of support and confidence, weighted confidence (WCF) of a rule $X \rightarrow Y$ can be calculated as the ratio of weighted support of ($X \rightarrow Y$) and the weighted support of X [9].For example, the weighted confidence of the rule R (Diabetes ="yes" → CVDs_label="high") can be calculated as:

$$WCF(R) = \frac{\begin{array}{c}\textit{sum of Tuple Weight having the condition}\\ \textit{Diabetes ="yes" and also given CVDs\_label = "high\_risk"}\end{array}}{\begin{array}{c}\textit{sum of Tuple Weight having the condition}\\ \textit{Diabetes ="yes"}\end{array}} \quad (2).$$

IV. EARLY WARNING MECHANISM BASED ON FUZZY LOGIC

In the decision making unit, some warning mechanisms or service models are built to decide what feedbacks should be given to patients through a reverse link. That's the reason why the early warning mechanisms are put forward. Early warning mechanisms are established by combining the rule set from the classifier and historical context information. We expect that a warning report will be forwarded to the user or the doctor when the verdict of illness is in a critical condition. What is more, different medical or dietary measures will be taken according to the different risk level. There are two types of early warning mechanisms to be discussed, the instantaneous risk early warning (short for IREW Mechanism) and long-term statistical early warning (short for LSEW Mechanism).

A. *IREW Mechanism*

The IREW mechanism is mainly targeting at a medical emergency such as sudden high blood pressure or a heart attack. According to the doctor's advice, a proper risk threshold is set for each parameter. Once the physiological parameter data come from the user domain, they will be compared with the risk thresholds. If they are much larger than the thresholds, it means that there is a serious possibility that the danger may happen. At this point, the decision making unit of the monitoring system should inform the family or a doctor of the danger.

This principle is very simple, but there is a tricky problem. That's "sharp boundary" problem about the continuous attribute. For example, if the normal range of heart rate is between 80-90 times/min, it does not make sense that a heart rate of 79 or 91 times/min is abnormal [11]. So, the adoption of fuzzy model is to establish accurate warning mechanism. The fuzzy process generally can be separated into three processes.

1) *Fuzzification*

The first process is fuzzification, which is to map crisp inputs to fuzzy variables by acquiring membership function value of the crisp inputs. Membership functions can take diverse shapes, and we use the simplest trapezoidal functions. The membership function for SBP is shown in Fig.4. Besides SBP parameters, heart rate, DBP and other parameters are also fuzzified.

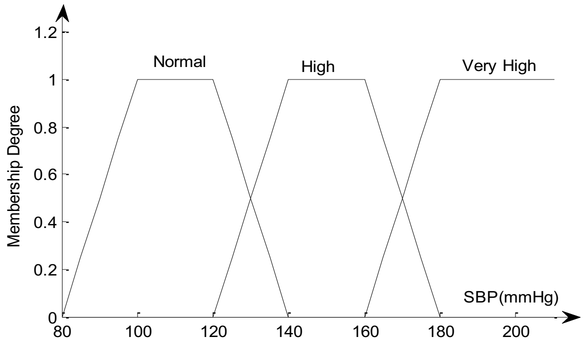

Fig.4.Membership functions for SBP

2) *Inference*

The second process is the inference based on some rules. Here, we adopt the rules with the form of ‘if-then’, since this form match the decision making process for healthcare monitoring. According to known rules, related fuzzy variables are used to infer a result.

3) *Defuzzification*

Generated fuzzy results may be not applicable for any applications. Hence converting the fuzzy variables into crisp outputs for further processing is sometimes essential, which is defuzzification. In our system, we adopt a typical and most widely used method, the centroid method. The output of this model is a crisp value, and then the value is compared with thresholds, shown in Fig.4, to make the comprehensive final decision on the patient status.

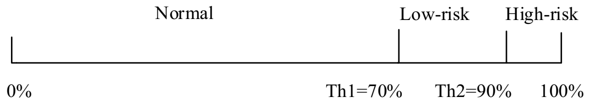

Fig.5.Thresholds of fuzzy model for healthcare monitoring

B. *LSEW Mechanism*

The LSEW mechanism mainly uses Rule Set (To be discussed in detail in Section V) from classifier to classify the CVD risk levels. It is worth noting that long-term statistical early warning is not real-time start-up and real-time feedback. It will set the unique monitoring cycle for each user according to his or her historical classification information. For instance, the warning cycle is set to a half year for a low risk patient, but three months for a high risk patient. Patients with different levels of risk will be given different prevention strategies. Moderate risk patients should be in close observation of BP, risk factors and target organ damage to decide when to use drug therapy. For patients at low risk, long-term observation, frequent BP checking and necessary intervention are needed.

V. Results and Analysis

This section shows an empirical performance evaluation of algorithm WAC, along with the well-known CBA and decision tree algorithm. Three chronic disease data sets are used, which are collected from several community hospitals of Haidian District. The experiments are composed of three parts. The first part is to compare WAC with CBA and C4.5 on accuracy. In the second part, we compare WAC with CBA and C4.5 in terms of the number of rules produced. What’s more, generated classification rules will be described.

A. *Classification Accuracy Comparison*

The accuracy is obtained by holdout approach [12], where 50% of the data are randomly chosen from the data set and used as training data and remaining 50% data are used as the testing data. The training data are utilized to construct a model for classification. After constructing the classifier, the test data are used to estimate the classifier performance. For WAC, we set the threshold of support to 10%, and for the confidence, its threshold is set to 85%. For CBA, all parameters have their default values. Discretization of quantitative domains was done using the entropy method [13].

TABLE VII. Classification Accuracy comparison

Data Set	WAC	CBA	C4.5
No.1	80.33%	79.42%	77.32%
No.2	83.15%	75.26%	75.83%
No.3	90.42%	93.7%	88.41%
Average Accuracy	84.63%	82.79%	80.52%

Classification Accuracy Comparison tabulates on the Table VII. Experimental results show that associative classifier is, in general, more accurate than that produced by classical decision tree C4.5. In addition, the average accuracy of WAC is higher than CBA. The main reason for better accuracy is that WAC assigns corresponding weight for each attribute according to priori domain knowledge. During the prediction, the attributes having more prediction capability can make more contribution. So only the weights of attributes are assigned reasonably, the average classification accuracy of WAC can be better than the general CBA.

A. *Number of Rules Comparison*

Using different algorithms will generate different number of rules for CVDs prediction. In this experiment, we compare the number of rules generated by C4.5, CBA and WAC. Fig.5

displays the comparative consequences. Clearly, WAC and CBA generated more rules than C4.5. It proves that Associative Classifiers (AC) can discover more rules than the traditional classification algorithm. While WAC generated fewer rules than CBA, meaning better understandability, even more stable performance.

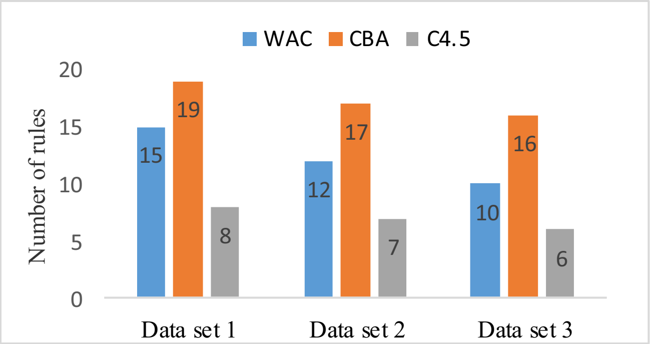

Fig.5. Number of rules comparison

C. *Classification Rule Sets*

From the Associative Classifiers built above, classification rule sets can be generated. The part of the rules are as follows:

Rules for high CVDs risk – contain 8 rules
Rule 1:
If SBP>=139 and Diabetes in ["no"] and Dyslipidemia in ["yes"] and Smoke in ["yes"]
then High risk;
...
Rules for moderate CVDs risk – contain 4 rules
Rule 1:
If SBP<139 and Diabetes in ["no"] and Dyslipidemia in ["yes"] and Smoke in ["no"] and Age>=55
then Moderate risk;
...
Rules for low CVDs risk – contain 3 rules
Rule 1:
If SBP <139 and DBP<89 and Diabetes in ["yes"] and Dyslipidemia in ["no"] and Smoke in ["no"] and Age <55
then Low risk;
...
Default: not clear

VI. Conclusions And Future Work

In the paper, we mainly discussed techniques in the prediction unit and decision making unit of CVDs monitoring system, i.e. CVDs risk classification algorithm and intelligent early warning mechanisms. On the basis of authoritative medical guidelines, we designed prediction workflow and used WAC to classify the risk level of CVDs. Experiments results indicate that WAC is more suitable for CVDs prediction with better accuracy and understandability comparing with C4.5 and CBA. Beyond that, intelligent early warning mechanisms were also designed. The instantaneous risk early warning mechanisms based on fuzzy logic technology can timely monitor physiological parameters mutation, which can prevent the occurrence of acute disease; Long-term statistical early warning mechanisms based on WAC classification rules can help patients to control chronic disease and to help doctors with diagnosis and clinical supervision. So, our study is of definite significance to help control risk level of CVDs patients.

So far, the article also has a few shortcomings. For example considering the space limit, the detailed theoretical analysis and experimental results of the warning mechanisms are not given. But we will use it as the focal point of the next work to perfect the service-oriented architecture of CVDs monitoring system.

In the future, we will consider that several kinds of advanced algorithms such as fuzzy logic, genetic algorithm can be combined with ACs to give more accurate results and fulfil the real life requirements.

References

[1] Yu Wang, GH Yang, JiaQi Ma, Death monitoring data of the National Disease SurveillanceSystem[M].Beijing: Military medical science press, pp.350-356,2010.

[2] Kang G. Wireless eHealth (WeHealth)—From concept to practice[C]//e-Health Networking, Applications and Services (Healthcom), 2012 IEEE 14th International Conference on. IEEE, 2012: 375-378.

[3] Di Lin; Xidong Zhang; Labeau, F.; Guixia Kang; , "A hypertension monitoring system and its system accuracy evaluation," e-Health Networking, Applications and Services (Healthcom), 2012 IEEE 14th International Conference on , vol., no., pp.132-137, 10-13 Oct. 2012.

[4] Wang N, Kang G. A monitoring system for type 2 diabetes mellitus[C]//e-Health Networking, Applications and Services (Healthcom), 2012 IEEE 14th International Conference on. IEEE, 2012: 62-67.

[5] Chinese Society of Cardiology of Chinese Medical Association, Editorial Board of Chinese Journal of Cardiology, Chinese Guide lines for Cardiovascular Disease Prevention, Chinese Journal of Cardiovascular Disease,vol.39,No.1,2011, 39(1).

[6] Writing Group of 2010 Chinese Guidelines for the Management of Hypertension, Chinese Guidelines for Management of Hypertension(2010), Chinese Journal of Cardiovascular Disease,vol.39,No.7,2011,pp.701-708.

[7] Ma B L W H Y. Integrating classification and association rule mining[C]//Proceedings of the 4th. 1998.

[8] Soni S, Vyas O P. Using associative classifiers for predictive analysis in health care data mining[J]. International Journal of Computer Applications, 2010, 4(5): 33-37.

[9] Soni S, Pillai J, Vyas O P. An associative classifier using weighted association rule[C]//Nature & Biologically Inspired Computing, 2009. NaBIC 2009. World Congress on. IEEE, 2009: 1492-1496.

[10] Soni J, Ansari U, Sharma D, et al. Intelligent and effective heart disease prediction system using weighted associative classifiers[J]. International Journal on Computer Science and Engineering, 2011, 3(6): 2385-2392.

[11] Agarwal S, Joshi A, Finin T, et al. A pervasive computing system for the operating room of the future[J]. Mobile Networks and Applications, 2007, 12(2-3): 215-228.

[12] Han, J., Kamber, M.: Data Mining: Concepts and Techniques.Morgan Kaufmann Publishers, Newyork (2001) [12] Han, J., Kamber, M.: Data Mining: Concepts and Techniques.Morgan Kaufmann Publishers, Newyork (2001)

[13] U. M. Fayyad and K. B. Irani, Multi-interval discretization of continuous-valued attributes for classification learning, in Proc. of the 13th Int. Joint Conf. on Artificial Intelligence (1993), pp. 1022-1027.