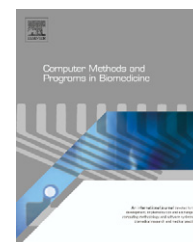




ELSEVIER

journal homepage: www.intl.elsevierhealth.com/journals/cmpb

A novel data mining mechanism considering bio-signal and environmental data with applications on asthma monitoring[☆]

Chao-Hui Lee^a, Jessie Chia-Yu Chen^a, Vincent S. Tseng^{a,b,*}

^a Department of Computer Science and Information Engineering, National Cheng Kung University, Tainan, Taiwan, ROC

^b Institute of Medical Informatics, National Cheng Kung University, Tainan, Taiwan, ROC

ARTICLE INFO

Article history:

Received 29 August 2009

Received in revised form

24 February 2010

Accepted 29 April 2010

Keywords:

Data mining

Bio-signal analysis

Environmental factors

Patient monitoring

Asthma attacks

ABSTRACT

Chronic asthmatic sufferers need to be constantly observed to prevent sudden attacks. In order to improve the efficiency and effectiveness of patient monitoring, we proposed in this paper a novel data mining mechanism for predicting attacks of chronic diseases by considering of both bio-signals of patients and environmental factors. We proposed two data mining methods, namely *Pattern Based Decision Tree (PBDT)* and *Pattern Based Class-Association Rule (PBCAR)*. Both methods integrate the concepts of sequential pattern mining to extract features of asthma attacks, and then build classifiers with the concepts of decision tree mining and rule-based method respectively. Besides the general clinical data of patients, we considered environmental factors, which are related to many chronic diseases. For experimental evaluations, we adopted the children asthma allergic dataset collated from a hospital in Taiwan as well as the environmental factors like weather and air pollutant data. The experimental results show that PBCAR delivers 86.89% of accuracy and 84.12% of recall, and PBDT shows 87.52% accuracy and 85.59 of recall. These results also indicate that our methods can perform high accuracy and recall on predictions of chronic disease attacks. The readable rules of both classifiers can provide patients and healthcare workers with insights on essential illness related information. At the same time, additional environmental factors of input data are also proven to be valuable in predicting attacks.

© 2010 Published by Elsevier Ireland Ltd.

1. Introduction

Chronic diseases are major illnesses and increasingly becoming part modern life. Irregular lifestyles, malnutrition, congenital genetic problems, are some of the major causes

of chronic diseases. Medical technology has advanced expeditiously in recent years, but the prediction of chronic disease attacks is still arduous. Therefore, patient monitoring is a necessary treatment for these patients. However, with the increasing number of chronic cases, medical human resources are insufficient. In recent years, many medical institutions

[☆] This paper is an extended version of the paper entitled "An Integrated Data Mining System for Patient Monitoring with Applications on Asthma Care" [20], by Vincent S. Tseng, Chao-Hui Lee, Jessie Chia-Yu Chen, which appeared in Proceedings of IEEE International Symposium on Computer-Based Medical Systems (CBMS), Finland, June 17–19, 2008.

* Corresponding author at: Department of Computer Science and Information Engineering, National Cheng Kung University, Tainan 701, Taiwan, ROC. Tel.: +886 6 2757575x62536; fax: +886 6 2747076.

E-mail address: tsengsm@mail.ncku.edu.tw (V.S. Tseng).

URL: <http://idb.csie.ncku.edu.tw/tsengsm> (V.S. Tseng).

0169-2607/\$ – see front matter © 2010 Published by Elsevier Ireland Ltd.

doi:[10.1016/j.cmpb.2010.04.016](https://doi.org/10.1016/j.cmpb.2010.04.016)

have tried to solve this problem by digitizing their medical equipments. Many physiology detection instruments are applied in patient monitoring. Doctors and healthcare workers can give a diagnosis based on the information detected by this equipment. Therefore, many chronic patients can use the tools to keep their data for advanced treatment. However, this mechanism is not sensitive for detecting sudden attacks of chronic diseases. Patients understand their conditions only after returning for medical treatment. For some chronic diseases which could cause fatal attacks, this mechanism is miserably insufficient.

With the prevalence of the Internet, this problem was solved rapidly. Many studies linked patients and doctors with many kinds of Internet services. Through the Internet, patients upload their data or symptoms to a specific server, whereupon doctors can access and analyze their data from the server. These services allow doctors to take care of their patients more readily and can inform the patients of crucial knowledge to prevent repetition of further attacks through this kind of techniques.

However, this kind of systems still requires doctors or medical experts to constantly monitor and cannot give real-time responses when attacks occur. The motivation is to construct an automatic system, which can warn patients through alarms and give basic medical instructions. Several researchers in the past built patient monitoring systems to predict attacks but most of them only considered a single factor of a target disease. Under this type of considerations, pivotal information for diagnosis maybe lost and attacks could not be predicted and alert given effectively. We believe that each disease related data contains information about attacks and taking this into account we can ascertain that the more information that is available the more it makes analysis and mining more reliable.

In addition, many related studies built systems for chronic diseases using varying methods. Some of them predicted chronic disease attacks through a statistical method. Some analyzed environmental data and built a patient monitor-

ing system by training a neural network. These frameworks did not inform patients and healthcare workers about the patients' condition except for a disease attack prediction.

Just as has been posited in many disease studies, e.g., the focus of Stockes et al. [18] was based mainly on patient data. Numerous patient disease related physiological profiles were taken into consideration. The findings showed a connection between these data profiles and the disease attacks using mathematic models which would then be used to predict when there would be attacks. In [7], Cure assembled a multi-layer perceptron (MLP) model neural network as the prediction module partly using environmental factors. Both methods predicted attacks with different disease related data and worked effectively. In terms of the integrity of the information, both still separately considered the related factors for attacks. Their outputs provide only the prediction results without any substantial readable or beneficial information for patients or healthcare workers.

In this paper, we aim to propose a novel data mining mechanism with applications on patient monitoring system for chronic diseases like asthma. We take varieties of chronic disease related factors into consideration, including both environmental factors and patient bio-signals. This design enables the system to produce highly accurate predictions on chronic disease attacks. In addition to predicting attacks, it also offers patients and healthcare workers comprehensive formularization in preventing attacks. Fig. 1 demonstrates the architecture of our proposed mechanism, in which the environmental pollutant data, climatic and atmospheric data and all disease related data are collected through the Internet and telecom communication. Subsequently, the system for chronic patient monitoring is built by analyzing the collected data. Two methods are developed in this system, namely Pattern Based Decision Tree (PBDT) and Pattern Based Class-Association Rule (PBCAR). In both methods, the concept of sequential pattern mining is adopted for feature extraction as the first step. As the result, the classifications of both methods are built as classifiers with tree-based and rule-based mining methods

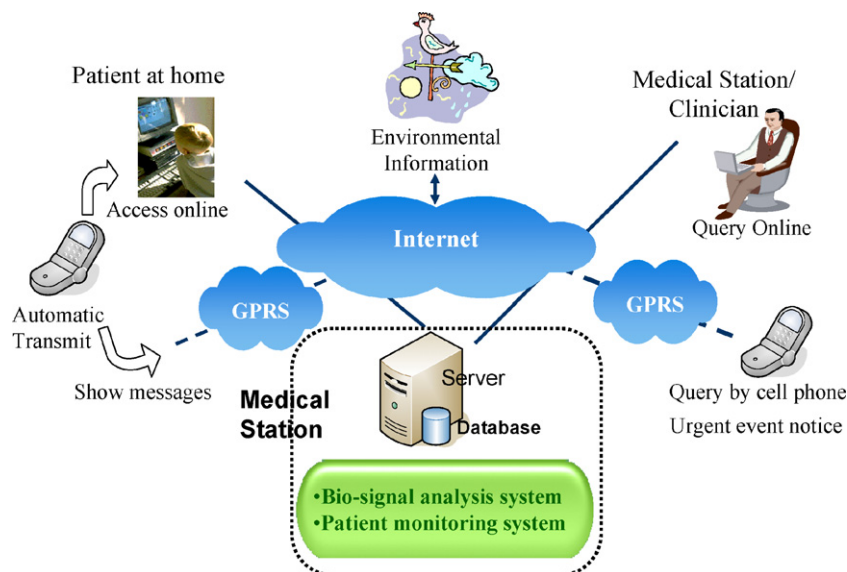


Fig. 1 – System architecture.

respectively. The classifier of PBDT illustrates each prediction result intuitively with only one tree path of feature set, so it is completely straightforward to appreciate the primary reasoning of a prediction result and its obvious suitability for easy diagnosis. The classifier of PBCAR shows all matched rules and the prediction results with its scoring method. The matched rules provide more facts about an illness and are appropriate for advanced studies by doctors and other medical experts.

In experimental evaluations, we appraised our system with an asthma related dataset, which consists of asthma allergic data, climatic and atmospheric data and the air pollutant data for the city where the patients reside. The patient clinical data were obtained from a hospital in southern Taiwan. The environmental data can be obtained from [24,25]. In experiments, both PBDT and PBCAR performed with high accuracy and high recall on this dataset. The matched rules of their classifiers not only perform predictions for asthma attacks but can also offer patients and healthcare workers the present conditions of patients. In addition, the experimental results show that the accuracy and the recall can be improved through the proposed system by taking into account more disease related factors, including bio-signals and environmental factors.

The rest of this paper is constructed as follows. The related works are discussed in Section 2. Section 3 mainly describes the proposed bio-signal mining system in detail. Subsequently, experimental evaluations are discussed in Section 4. Finally, we summarized the several key points of this research and proposed some possible future undertakings in Section 5.

2. Related works

Due to the rapid development of society and technology, medical services have become ever more universal. Vast numbers of patients can now be treated, but the increased pressure on existing medical resources has caused traditional medical services to be unable to cope efficiently with this influx of patients. Thus, digitalization for medical services is crucial. In [21], Várady et al. described that the process of digitalization can be regarded as three stages. In their first stage, the *Hospital Information System* (HIS) integrates all medical information in the hospital. Simultaneously, doctors and nurses can access all medical data through the *Patient Data Management System* (PDMS). In the second stage, a *Decision Support System* would be integrated with HIS to assist doctors to analyze patients' bio-signals. In recent years, the rapid development of technology has brought medical systems into the third stage. In this stage, the systems have integrated Internet technology and extended their service ranges to become the *Home-Care Monitoring System*.

Asthma is a common chronic disease, especially in children who cannot take care of themselves. The number of children with asthma in Taiwan grows by at least 1% per year. Many hospitals digitalized their asthma monitoring system to reduce their workload. This situation has initiated substantial asthma related research. Some research focused on a study concentrating on patients' DNA profile. O'Leary et al. [15] addressed the allergens of asthma. Through the lung functions of children, asthma was able to be detected using neural

network approaches [3]. Finkelstein et al. [8] further assembled a communication platform between asthma sufferers and emergency rooms for an advanced research.

In [10], Jan et al. constructed an Internet-based interactive tele-monitoring system for asthmatic children, named *Blue Angel Asthma Home-Care System*. Through that system, they expected to enhance the generalization, convenience, and efficiency of the home-care tool. According to the description of that research, healthcare workers would inform asthma patients of basic medical information, for example, important symptoms of asthma and the usage of emergency asthma drugs. At the same time, patients were informed on how to take care of themselves by using a peak expiratory flow meter (PEF). In addition, a network platform was also introduced. The patients could upload their daily monitoring records and obtain asthma related information through this tool as well. It was posited that the mechanism makes disease control more efficient and effective, and reduce the probability of asthma attacks.

Chu et al. [6] proposed a ubiquitous warning system for asthma-inducement, which took the Global Position System (GPS) into consideration. According to the data from the GPS, the system could obtain and display daily averaged temperatures, humidity levels, and atmospheric pressure from the Central Weather Bureau [24] and air pollution reports from the Environmental Protection Administrator Executive Yuan [25] to build a decision module by MLP neural network. Therefore, their system could assist healthcare workers to make a diagnosis and help to locate patients expediently through GPS information when asthma attacks occur. However, the system did not take a patient's individual information into consideration, and could not offer patients any related alarm to prevent attacks. In addition, applying statistical methodologies is also a popular method for processing medical data. In [18], Stockes et al. explored and undertook the task of discovering the importance of asthma bio-signals using statistical methods. As the result, they proposed the Entelos® Asthma PhysioLab™ System (EAPS), i.e., a module used for asthma patients, which considers molecular cells, organ, allergy symptoms, respiratory flow, and patients' drug reaction applying a mathematical function and principle; Through this system, they further found a relationship between the immune system and asthma attacks.

In the field of data mining, association rule mining [1] and sequential pattern mining [2] are the major methods for discovering the ubiquitous relationship in a dataset. Association rule mining ascertains frequent connections and places all item sets in a dataset. Under the same concept, sequential pattern mining undertakes to mine out the relationship between items or item sets in chronological order. They are then usually able to represent features of a dataset. On the other hand, decision tree mining [16,17] is an important data mining method for classification, while many researches have addressed this and applied this kind of approaches into many applications. However, the types of datasets have changed expeditiously in recent years as copious amounts of information need to be recorded. Therefore, many integration mining methods were proposed for handling this kind of datasets. In [13], classification was performed by the concept of association rule mining. It took the relationship between items as considered attributes

for classification and made their method more suitable for item-based data classification. For a similar purpose, Lesh et al. [12] applied the concept of sequential pattern processing to time series classification. Some advanced and involved techniques are also applied in medical related studies, such as the application in an expert system [5], and the classification mechanism on the electrocardiogram (ECG) data [19]. In this paper, we integrated several data mining techniques in order to demonstrate that our proposed method can provide an effective alert mechanism for attacks on chronic diseases like asthma by considering more critical factors, while further details are described as in the following section. This is an extended paper of our preliminary work in [20].

3. Integrated bio-signal data mining mechanism

In this research, we designed a mechanism to predict asthma attacks automatically by mining asthma physiological symptoms and environmental factors together. We assume that there are useful features about asthma symptoms in these data. Thus, we used data mining methods to discover features of asthma attacks and construct a classification mechanism for those features in order to establish an alarm system and to aid healthcare practitioners. For most patients with chronic diseases who need to be monitored, hospitals usually record their daily disease related data. Doctors or healthcare practitioners can do long-term observations and advanced studies to find more suitable treatments for the patients whose data have been gathered. These recorded data consist of patients' clinical data and their bio-signals, which are contiguous values and could be regarded as time series data. In addition, the outbreaks of many chronic diseases are often associated with environmental factors. Asthma as an example is not only bronchial contraction but also a series of inflammatory reactions. This means that attacks can occur at anytime when a patient's airway is stimulated continuously by environmental allergens. Consequently, our research took into account most environmental factors. In the same concept of patients' daily records, we regarded the long-term records of environmental factors as time series datasets as well. In the following sub-sections, the problem definition is firstly described. After illustrated the proposed system architecture, the procedures of constructing Pattern Based Decision Tree and conducting the Pattern Based Class-Association Rules are presented.

3.1. Problem definition

For achieving our mechanism, two major challenges should be conquered. The first challenge is how to seek or design a suitable processing for these complex datasets which consist of many different-type contiguous records. The second one is to make our prediction mechanism of asthma attacks interpretable. Therefore, we did format transformation for different-type data to overcome these two challenges. Fig. 2 shows a basic architecture of our problem. Precisely speaking, we want to know if one patient will suffer from an asthma attack in near future, when all his asthma related information can be got, which contains his asthma allergic symptoms and

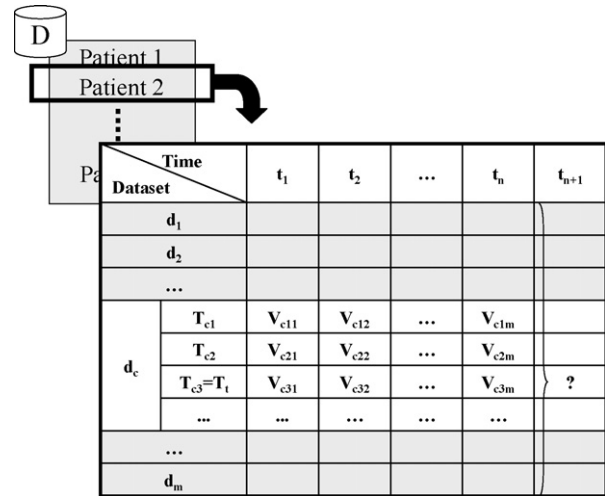


Fig. 2 – Illustration of one patient's data.

the relative environmental conditions. Under our assumption, all asthma patients' historic data could be granted as a knowledge base and these data can be integrated into a huge multivariate time series dataset, while each instance of the dataset is the data of one patient.

Based on all the descriptions above, we formulated our problem and defined it in the following manner. Consider a dataset D , in which all the related data of monitored patients are collected. The data of each patient contain all items afore mentioned. Thus, it composed by all related datasets $\{d_c | c$ is an integer and represents the index of related datasets in $D\}$, such as datasets of weather, air pollution, and patient allergy. For each related dataset d_c , it consists of time series data $\{T_{cr} | r$ must be an integer and represents the index of factors in $d_c\}$. T_{cr} represents the records of each factor in this dataset. These factors of a related dataset are generally recorded as a time series pattern, such as CO_2 or NO of air pollution dataset. Each factor T_{cr} is formed by a sequence of contiguous recorded values $\{V_{cr1}, V_{cr2}, \dots, V_{crm}\}$. Our purpose is to target some T_t in all T_{cr} s as an aim and to build a classifier C for T_t with the dataset D . In chronic disease dataset, D is the dataset containing all disease related datasets d_c , such as daily records of patients, weather condition and air pollutants. T_{cr} represents each factor of related datasets, such as concentration of CO_2 in air pollutant dataset. The T_t would be a disease state or a diagnosis result for a patient. In terms of asthma monitoring, T_t is the daily data which record whether the patient suffered an attack on that day. In our proposed methods, we firstly extracted classifiable sequence features $\{V_{q1}, V_{q2}, \dots, V_{qx} | 1 \leq x \leq m (cr1 \leq qx \leq crm) \text{ and } q(j-1) < qj \text{ when } 2 \leq j \leq x\}$ from all T_{cr} . Subsequently, we built a classification model for T_t using data mining methods with these features and then T_t could be predicted using the configured model.

3.2. System architecture

Fig. 3 shows the architecture of our proposed mechanism. Patients' bio-signal data and environmental data of patients' locations are considered in this mechanism. Through the preprocessing elements, these datasets are integrated and

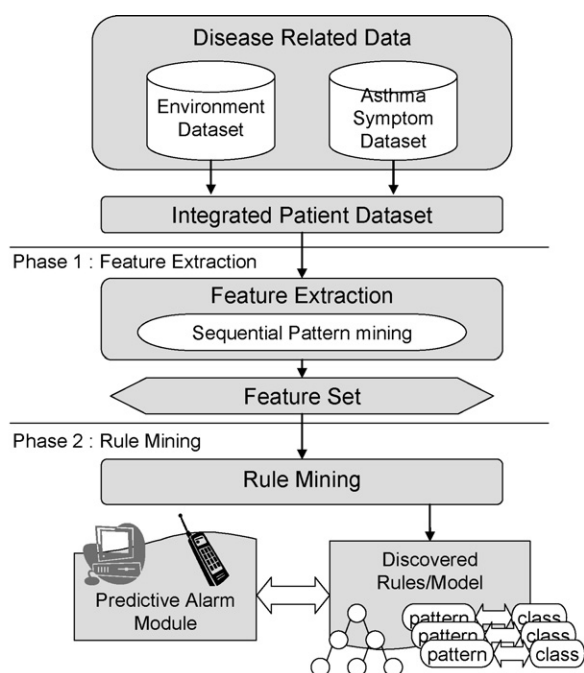


Fig. 3 – Work flow of the two-phase design for the proposed mechanism.

transformed into a specific dataset with medical and environmental knowledge for mining. Subsequently, two alternative methods were designed for the data mining segment. Finally, the disease related diagnosis rules would be discovered or the prediction model would be built. These results could aid healthcare practitioners in disease diagnosis and construct an advanced predictive alarm mechanism.

Considering the researches which are related to disease prediction and prevention, most of them have only focused on physiological symptom data. In General, nothing more than prediction results could be provided in their predictions. However, the attacks of many chronic diseases are very much related to environmental factors. Therefore, patients' bio-signal data and environmental factors are considered together in our mechanism. These data are comprised of daily physiological data, past disease attack records, daily climatic conditions, daily weather statuses and daily records of air pollutant concentrations. Substantial disease related information mining from these data makes the disease prediction be able to be built more completely. Moreover, our prediction model offers more thorough information and rationalization for the prediction results and can be used as a reference in diagnosis or further advanced studies.

Traditional predictive methods often cannot be applied to such a complex dataset considered in this paper. The dataset consists of time series data and attribute-based data, which are two different formats. The most values of bio-signal data and environmental data are of a numerical type. Consequently, we integrated the mining method for time series data to extract data features and the classification method for attribute-based data to build our mechanism of disease attack prediction.

In order to reduce problem complexity, and to increase the medical significance and readability of outputs, we symbolized these numerical values with the meaning and information of each value on its related studies. For example, the concentration of environmental factor CO is a numerical attribute, but we can symbolize its value into different levels of air pollution. This discretization method of numerical value is more convenient to be conducted and makes the result of any analysis more comprehensive. Resultantly, we reviewed all related studies of each data and chose appropriate policies for data preprocessing, such as common level definitions for some air pollutants. In our proposed mechanism if there was no suitable expertise of one factor data with numerical, some common methods of data mining could be applied directly, such as piecewise aggregate approximation (PAA).

In addition, feature selection can be considered according to input datasets. In our proposed mechanism, the purpose of feature extraction is mainly to derive characteristics from multivariate time series datasets. All extracted features could be meaningful and applied directly on the following classification mining. However, feature selection also can be considered to be performed after feature extraction under some conditions, such as a huge quantity of input dataset or space limitation for the final classifier. We suggest choosing the feature selection methods which do not destroy or rebuild features, such as relief [11], information gain (IG) [9], or mutual information (MI) [4].

As shown in Fig. 3, we adopted a two-phase concept to perform the integrated data mining. The concept consists of a feature extraction phase and a rule mining phase. After preprocessing, there could still be differences in the format of data. We extracted all disease related features in the feature extraction phase. Then, all the features could be rearranged in some specified format. Finally, the rules or the model for disease attack prediction can be constructed by the features in the last part of the mechanism. Based on the proposed architecture, we presented two policies for asthma attack prediction. The first policy is named *Pattern Based Decision Tree* (PBDT), which integrates sequential pattern mining and decision tree mining; the second is named *Pattern Based Class-Association Rule* (PBCAR), which merges sequential pattern mining and classification-by-associations mining.

3.3. PBDT: Pattern Based Decision Tree

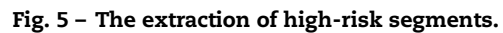
We designed a tree-based method, named the PBDT, for asthma high-risk prediction, which follows the two-phase design and the architecture shown in Fig. 3. After data preprocessing, the concept of sequential pattern mining is adopted to retrieve the representative sequences of these time series data as static features in the first phase. Secondly, the decision tree is assembled with these extracted features. The pseudo-code of the PBDT algorithm is shown in Fig. 4 and its implementation on all asthma related dataset is described in the following sub-sections.

3.3.1. Phase 1 of PBDT: feature extraction

Most prepared related data are time series datasets, such as asthma symptom dataset and environmental dataset. These types of datasets are more difficult to comprehend for users

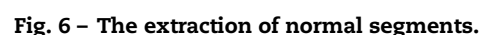
Fig. 4 – PBDT algorithm.

In the implementation of this phase of the PBDT, a parameter λ is defined as *observation interval* and it represents the number of the days for observation. The value of λ must be evaluated by experiments or suggested by healthcare practitioners. Through the patients' clinical data and diagnostic records, each date of asthma attacks can be recognized and the data within λ day before an asthma attack are cut as a high-risk segment. Each dataset, i.e. asthma dataset, weather dataset, and pollution dataset, must be treated with the same extraction method so as to get data intervals that contain high-risk features. Fig. 5 shows that how to cut high-risk segments of asthma attacks of one patient's record on asthma dataset with $\lambda = 5$. The sequence in Fig. 5 is one symptom in asthma allergic data. The symptom is recorded as “g”, “y” or “r”, which mean the asthma risk light grades using in general



After above procedures, the datasets of high-risk and normal segments were prepared, and sequential pattern mining performed on these two segment sets in order to find their high frequency sub-sequences. Fig. 7 depicts a simple map of sequential pattern mining on normal segments, e.g., frequent sub-sequences, such as “g” and “gyg”, can be discovered through this process. These high frequency sub-sequences are asthma related features, and would follow the following format:

Most chronic related research focused on all accessible patients' bio-signals, yet many environmental factors which could affect chronic disease attacks are ignored. The major goal of this study is to predict asthma attacks. Therefore, the



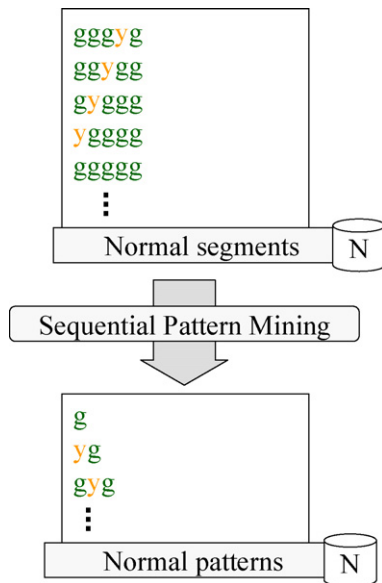


Fig. 7 – The feature extraction of normal segments.

PBDT takes all accessible environmental factors of a patient's residence and immediate surroundings into consideration. All environmental factor data are numerical, and the PBDT can symbolize each value to simplify the following procedures in the preprocessing stage. Subsequently, the data should be cut into high-risk and normal segments following the setting of the parameter λ . In order to extract environmental factor segments corresponding to the asthma related features, all time intervals of high-risk and normal segments of each patient in asthma dataset must be collected firstly. Then, all data in these time intervals of high-risk segments are extracted as high-risk segments for each environmental factor dataset and normal segments are also obtained using the same processing manner. Finally, two segment sets are prepared respectively for each environmental factor. There are numerous environmental factors, such as daily averaged temperature, and the daily concentrations of carbon dioxide, so we can get high-risk and normal datasets for each environmental factor. Fig. 8 shows a high-risk segment of a daily averaged temperature dataset is extracted by matching the time interval of a high-risk segment of the asthma dataset. The value sequence {28.4, 28.4, 27.8, 28, 25.3} which represents the temperature situations can be accessed before this patient suffer from the asthma attack.

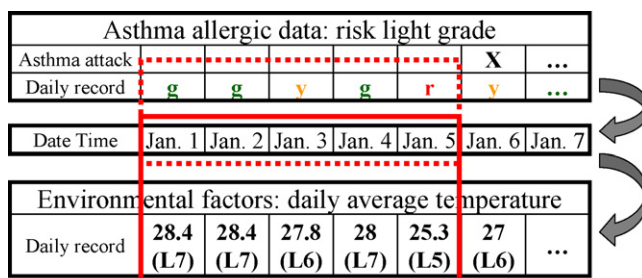


Fig. 8 – Extraction of a high-risk segment from daily temperature data.

Table 1 – The time intervals of all segments.

	Time interval	High risk?
Segment 1	patient 1: $d_{11}, d_{12}, \dots, d_{1\lambda}$	Y
$\vdots$	$\vdots$	$\vdots$
Segment k	patient e: $d_{k1}, d_{k2}, \dots, d_{k\lambda}$	Y
Segment (j + 1)	patient 1: $d_{(j+h)1}, d_{(j+h)2}, \dots, d_{(j+h)\lambda}$	N
$\vdots$	$\vdots$	$\vdots$
Segment (j + h)	patient e: $d_{(j+h)1}, d_{(j+h)2}, \dots, d_{(j+h)\lambda}$	N

According to the design of phase 1, the PBDT performs sequential pattern mining to discover features of each environmental factor on its high-risk and normal segment sets. Thus, high frequent sub-sequences of high-risk segments and normal segments for each environmental factor can be extracted and have the following format:

Pattern P_e from high-risk segments in environmental factor $M:\{\text{state } i, \text{state } j\}$.

Pattern N_e from normal segments in environmental factor $M:\{\text{state } x, \text{state } y, \text{state } z\}$.

All features of each asthma related dataset were prepared, which contain patient clinical data, asthma symptom dataset, weather factor dataset, and air pollution dataset. These features are integrated as an indexed table for the second phase of the PBDT. Note that the two important data tables must be generated. The first table contained the time intervals of high-risk segments and normal segments. All high-risk and normal segments of patients must be identified by these time intervals. Table 1 presents a sketch of this table. Another table listed all features extracted with sequential pattern mining respectively from high-risk and normal segments of each sub-datasets. The sketch of this table is shown in Table 2.

3.3.2. Phase 2 of PBDT: rule mining

In this phase, a disease classifier was built with all the extracted information from the last phase. After phase 1 of the PBDT, the all asthma related high-risk and normal frequent features were prepared. Firstly, the high-risk and normal seg-

Table 2 – All features of all asthma related factors.

Dataset	Factors	The features of high-risk segments and normal segments
Asthma symptom data	Allergic (1-n)	Pattern P_{a1}
		$\vdots$
Environmental data	Factor 1 (1-q)	Pattern P_{an}
		$\vdots$
		Pattern P_{e11}
		$\vdots$
	$\vdots$	Pattern P_{e1q}
		$\vdots$
	Factor m (1-r)	$\vdots$
		Pattern P_{emr}

Table 3 – An attribute table of all segments.

	Factor X_1			...		Factor X_m	High risk?
	Pattern 1	...	Pattern n_1	...	...	Pattern n_m	
Segment 1	0	...	0	...	...	1	Y
...	...	...	...	...	...	...	...
Segment j	1	...	1	...	...	0	Y
Segment $(j+1)$	0	...	1	...	...	0	N
...	...	...	...	...	...	...	...
Segment $(j+h)$	1	...	1	...	...	1	N

ments of each factor data were transformed into the attribute table with their own frequency features. This processing was represented from lines 8 to 18 in the PBDT algorithm and makes traditional decision tree mining performing on the datasets having time series data. Subsequently, as the last line of the PBDT algorithm in Fig. 4, decision tree mining is performed to get a classification model. Its tree-based model can predict asthma attacks and the features related to asthma attacks can be informed from its matching paths. In practice, the process consists of the following two steps:

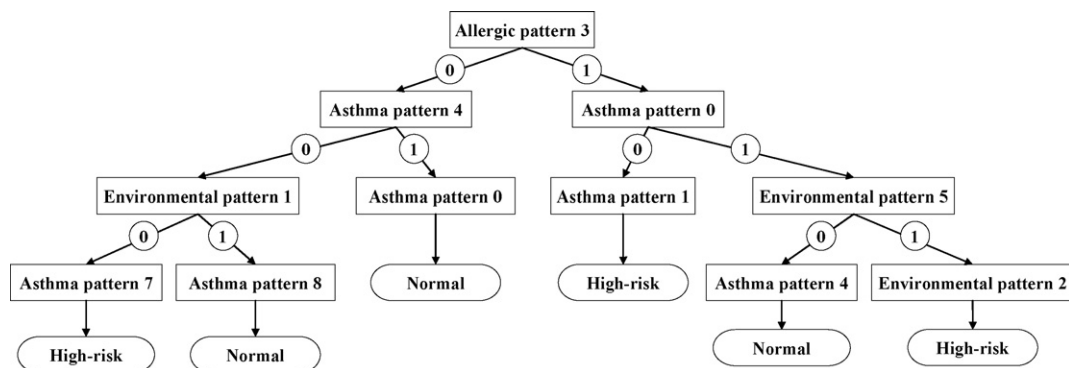
1. In the first step, a classification module is built with decision tree mining. The tables prepared in phase 1 of the PBDT must be integrated and transformed into an attribute table. Each segment contains many sequence data of different asthma related factors and it is processed as a transaction in the attribute table. All features of all asthma related factors are attributes. The table keeps all connections between transactions and attributes. So that all categories, and types of features and transactions can be accessed from this table. Table 3 presents the sketch of the attribute table. The grid of segment p and factor pattern q is 1 when all data in the time interval of segment p contain the pattern q . Otherwise, it should be 0. The grids of high-risk attribute follow the high-risk values of segments.
2. After the attribute table is established, all relations between features, high-risk segments and normal segments are labeled. Therefore, decision tree mining can be performed on this table with setting “High risk?” attribute as target to generate a decision tree. The decision tree is a binary tree and like the tree-like structure shown in Fig. 9. Each tree node (i.e., leaf) of the tree represents one feature, and two

sub-trees of a tree node are under the conditions with the feature and without the feature respectively. All paths in the tree are oriented to produce one result, so a prediction can be made if an unknown segment is high-risk by tracing the tree with its feature set. Fig. 9 shows a sketch of the tree with more details.

According to the PBDT classifier, after a patient completes his daily routine records, the system obtains all related data in last λ days, which contain asthma allergic symptoms and environmental factors. Through the matching on sequential patterns, the feature set of these related data could be extracted. Subsequently, the “High risk” result of next day can be predicted by tracing the tree with the specific feature set. For each reached node, its sub-tree matched by the condition of the feature set is chosen for the rest tracing. The tracing stops at one leaf node and the label of the leaf node is the prediction result. Fig. 10 presents the flow chart of the prediction method of our proposed PBDT algorithm.

3.4. PBCAR: Pattern Based Class-Association Rule

In the concept for the PBDT, asthma attacks are predicted using traditional decision tree mining. However, in [22] it was described that the rule set of a decision tree could be covered by the rule set of the association-based method, such as the Classification Based on Associations (CBA) [13]. The reason being the decision tree mining adopts greedy algorithm as its policy. Several rules must be omitted for the process of tree building. Therefore, we enhanced the algorithm of PBDT, and proposed a new method which integrated the concept of sequential pattern mining and association-based classifier,

**Fig. 9 – Example of decision tree for asthma attack classification.**

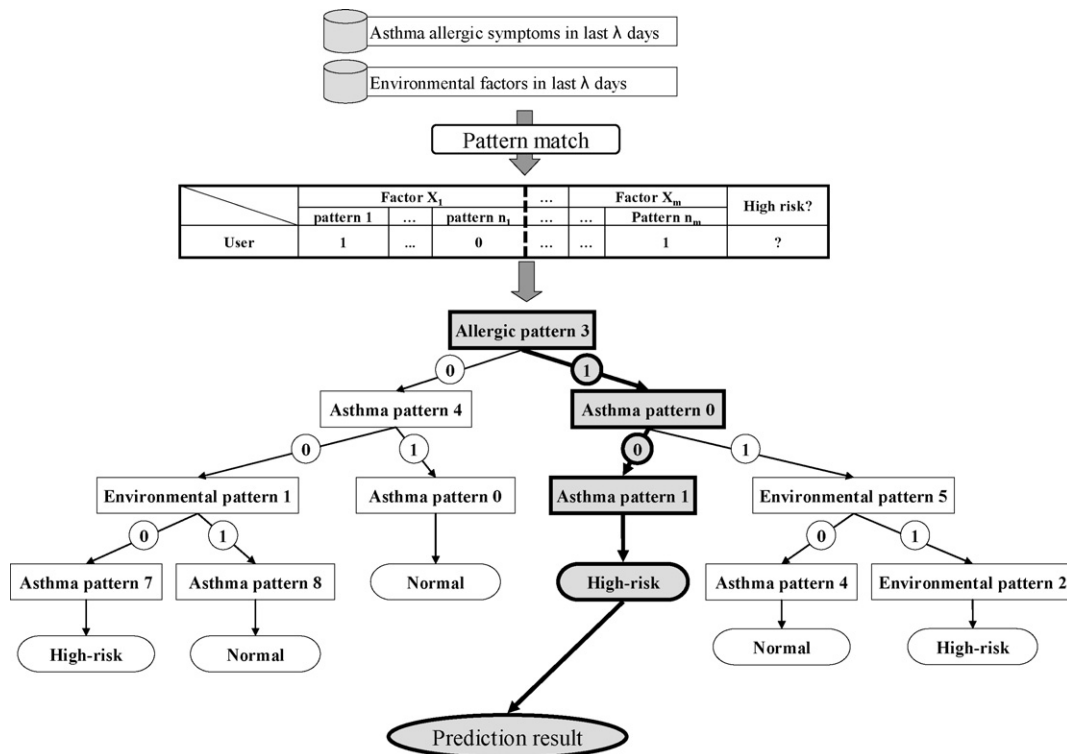


Fig. 10 – Asthma attack prediction with PBDT.

named Pattern Based Class-Association Rule (PBCAR). The system architecture of the PBCAR is based on the architecture of Fig. 3 and the relative algorithm is shown in Fig. 11. Two major phases of PBCAR are described in detail in the following segments.

3.4.1. Phase 1 of PBCAR: feature extraction

In phase 1 of the PBCAR, the same processing of phase 1 of the PBDT is performed. In practice, the same two tables also can be acquired as is mentioned in phase 1 of the PBCAR. Therefore, phase 1 of the algorithm PBCAR is the same as phase 1 of the PBDT algorithm.

3.4.2. Phase 2 of PBCAR: rule mining

In this phase, a classifier was built by the concept of the CBA in the PBCAR algorithm. Therefore, our PBCAR algorithm transforms the whole dataset into a transaction dataset with the feature set F from lines 8 to 17 in Fig. 11. The CBA-like algorithm designed and shown from lines 18 to 24 can be performed on the transaction dataset directly. The support and confidence for each class of all generated item sets should be counted as well. Subsequently, the redundant items can be filtered out with a support and confidence threshold setting. The rest of the item sets are collected as an association rule set, in which each rule has a score for its own class. Finally, the asthma attack prediction can be performed using these scored association rules.

In practice, after all related data were prepared in phase 1 of the PBCAR a classifier was built with a CBA-like method in this phase. There are several factors which need to be

defined before the method, such as thresholds, scoring function and scoring method for classification. We explain the whole method in the following four steps.

- Two tables of phase 1 should be integrated into a transaction table in this step. Table 4 shows the result, in which the features and class label of each segment are listed clearly. In this table, each segment is a transaction, and each feature is an item of the segment. In order to fill the table, all features of all factors must be checked if they are contained for each segment. Each data segment consists of data sequences of different factors. Each data sequence may contain several features of its factor. Subsequently, the table collects all contained features of all factors as the item set for each segment. After collection of item sets of all transactions, the high-risk attribute of each segment is appended to this transaction table.

Table 4 – A transaction based data table.

	Features {Factor, Pattern}	High risk?
Segment 1	$\{X_1, P_1\}, \{X_1, P_3\}, \dots, \{X_m, P_4\}$	Y
...	...	...
Segment j	$\{X_2, P_3\}, \{X_3, P_1\}, \dots, \{X_{m-2}, P_1\}$	Y
Segment (j + 1)	$\{X_6, P_2\}, \{X_1, P_3\}, \dots, \{X_{m-1}, P_3\}$	N
...	...	...
Segment (j + k)	$\{X_3, P_1\}, \{X_5, P_3\}, \dots, \{X_m, P_5\}$	N

```

Algorithm PBCAR
Input: A dataset consists of considered factors ( $D = \{d_c | c > 0\}$ ),
      Observation interval ( $\lambda$ ),
      Minimum support (min_sup)
Output: a rule-based disease classifier (PBCAR_classifier)
Procedure PBCAR ( $D, w, \text{min\_sup}$ )
Phase 1: //Feature Extraction
1  let  $D_p$  be the dataset which is rearranged data  $D$  with patient id.
2   $t_a = \text{ExtractDiseaseAttacks}(D_p)$ ; //To collect time points of disease attacks
3   $S_a = \text{HighRiskSegmentsExtraction}(D, t_a, \lambda)$ ;
4   $S_n = \text{NormalSegmentsExtraction}(D - S_a, \lambda)$ ;
5   $F_a = \text{FeatureExtraction}(S_a)$ ;
6   $F_n = \text{FeatureExtraction}(S_n)$ ;
7   $F = F_a + F_n$ ;
Phase 2: //The table of features  $F$  is prepared after phase 1
8  set a transaction dataset  $T = \emptyset$ 
9  for each patient data  $d_p \in D_p$ 
10   set a transaction  $n = \emptyset$ 
11   for each feature  $f \in F$ 
12    if  $d_p$  contain  $f$ 
13      $n.add(f)$ ;
14   end for
15    $n.class = d_p.class$ 
16    $T = T \cup n$ 
17 end for
18  $\text{CBR\_set} = \text{associationMiningWithClass}(T)$ ;
19 for each itemset  $r \in \text{CAR\_set}$ 
20    $r.score = \{(r\_class\_label) \rightarrow class\_label\}.confidence$ ;
21 end for
22  $k = \text{evaluateScoreBalance}(\text{CAR\_set})$ ;
23  $\text{classifier} = \text{buildingScorePolicy}(\text{CAR\_set}, k)$ ;
24 return classifier;

```

Fig. 11 – PBCAR algorithm.

- In this step, the classifiable frequent item sets are extracted. Therefore, association rule mining was adopted here to extract frequent item sets firstly. After frequent item sets are collected, the confidence of item sets for each class is counted. Finally, the classifiable frequent item sets are prepared in which each one is frequent and keeps confidences of each class. The CBA method [13] can predict the class of a novel transaction with these item sets. However, the classification policy of the CBA method is too simple to have a good accuracy, we then further modify the classification method in next step.
- The purpose of this step is to build a scoring mechanism. Scoring methods are widely applied in rule-based classification methods. Basically, the CBA classify a new data by the matched class-association rule which have the highest confidence of the result class. Though the method can produce the result rapidly, the considered factors are not enough. In [14], Liu et al. enhanced the CBA and proposed a scoring method, which performs classification by considering all rules. Following their concept, each rule keeps different scores for different classes. For each rule, the score of a class is defined as the value which is the confidence of that class multiplied by the support. Thus, a new case

The score of each matched rule of both classes (Positive/Negative)

- Positive Score = support * confidence
- Negative Score = support * confidence / k
- k is a constant value to adjust the imbalance scores.

Classification

```

If (Positive score > negative score) result = positive
else result = negative
return result

```

Fig. 12 – Scoring function of class-association rules of the PBCAR.

is classified with the cumulated scores of all classes from matched rules in the PBCAR.

- The scoring method of the PBCAR takes more factors into consideration, but there are still some irregularities. Most datasets for classification have imbalanced distribution of the target attribute, especially disease prediction datasets. The asthma dataset also has this problem, that is, there are fewer high-risk (positive) cases than normal (negative) ones. The skew datasets make a huge difference between the quantity of positive features and negative features. That promotes the results of the scoring method to be more negative, and reduces the recall of positive simultaneously. Therefore, we referred the work in [14] to define a parameter k as *balance parameter* to fix this imbalance in the two-class dataset. Fig. 12 shows the definition of modified score function in the PBCAR. All normal scores of rules are needed to be divided by k , so that the final cumulated scores of two classes are more appropriate for classification. The value of the parameter k generally needs to be formulated by experts or set by experimental evaluations. In regard of this, we did experiments to evaluate its value and discuss its effects, which are discussed and explained in detail Section 4.

After the four steps of building classifiers, the generated rule set appears as is represented in Fig. 13. According to the patient's daily record, the PBCAR gathers all data sequences of related factor datasets in λ days. The data sequences are transformed into a feature set like a transaction in Table 4 by matching the ones in Table 2. Then, the PBCAR checks the rule set of Fig. 13 and marks all matched rules. Finally, the PBCAR classifier accumulates the scores of normal and high-risk ones of these matched rules and gives the patient the result keeping higher score as a prediction result for the coming tomorrow.

If the second rule in Fig. 13 is marked, it represents the existence of asthma related data and high temperature differ-

```

{Photophobia, Apply controller} sup: 0.51, conf: [ N: 0.2, A: 0.8 ]
{High temperature difference, PM10 is high, Fever} sup: 0.45, conf: [ N: 0.3, A: 0.7 ]
{No treatment advised, No air pollution} sup: 0.33, conf: [ N: 0.68, A: 0.32 ]
...
{CO2 is high, PM10 is high} sup: 0.31, conf: [ N: 0.35, A: 0.65 ]
and balance parameter k = 2

```

Fig. 13 – Class-association rule of the proposed PBCAR for classification.

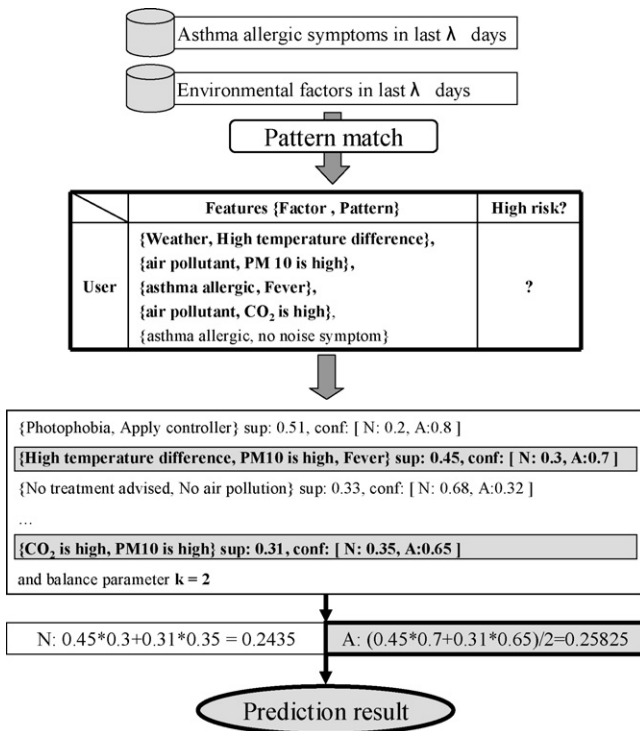


Fig. 14 – Prediction model of the PBCAR.

ence, a high PM10 value of air pollutant data indicates that the patient had a fever in the past λ days. So according to these features, if this rule keeps the highest confidence of all marked rules, the CBA predict that the patient will suffer from an asthma attack at $(\lambda + 1)$ th day. However, our PBCAR method takes the scores of all marked rules into consideration, Fig. 13 just offers a score $0.45 \times 0.3/2 = 0.0675$ for normal and a score of $0.45 \times 0.7 = 0.315$ for high-risk respectively.

These classification steps are shown as a flow chart in Fig. 14. There are two differences in classification method from the PBDT. (1) Input data must be transformed into transaction format instead of attribute format. (2) More rules are matched. Those differences make the characteristics of a PBCAR classifier fundamentally different from a PBDT classifier for users. In Fig. 14, patient's asthma related data in last λ days are transformed into the item set. The item set represents all matched features discovered in phase 1. These items may be allergic features or environmental ones. Subsequently, each rule generated by PBCAR must be checked whether all its items are contained by patient's feature set. Finally, the prediction result of the patient status is the class label with the highest cumulated score of contained rules.

Firstly both methods proposed are mining normal and high-risk sequential patterns as features. Then they build different classifiers through the concept of tree-based mining and rule-based mining respectively. As a result, the data format of both methods and the usage of both classifiers have different benefits. The PBDT referred the concept of decision tree mining. The concept tries to cover whole cases in training data with few major attributes in the tree architecture. Therefore, decision tree mining adopted a greedy algorithm to select a few rules for the tree building. This policy makes the

classification easy to perform and be comprehended but may lose considerable information for classification as well. In contrast, the tree architecture needs less capacity. It is suitable to be embedded in mobile equipments which have less memory but gives adequate and immediate alarms. In the method of our PBCAR, we adopted the concept of the CBA to build a class-association rule set. All possible features for classification are enumerated as rules. Therefore, when new data needs to be classified, all rules are considered. All matched rules determine the classification result by the scoring method. Although this policy makes the matched rules more complex, they can offer more detailed information of a classification, especially for healthcare workers.

4. Experimental results

4.1. Data

In this work, we adopted a real asthma dataset to evaluate our proposed architecture and methods. The real asthma dataset contains two sub-datasets. The first one is the dataset with the patients' asthma related data, which consists of asthma data and patients' diagnostic records. The second dataset includes several environmental factors, which contains weather data and air pollutant data. In addition, the first sub-dataset was collected by the Blue Angel 24 h Monitoring System of Tainan Allergic Asthma Infants Health Association [26]. The asthma data of the first sub-dataset contain the largest respiratory flow of the daily morning and evening, fever symptoms, daytime asthma symptoms, nighttime asthma symptoms, and the corresponding risk level of an asthma attack; while asthma symptoms include the allergic states of the eyes, nose and skin. The patients' diagnostic records comprise patient id, age, height, asthma medicine instructions and doctors' recommendations. Therefore, each daily record of the first sub-dataset consists of 23 attributes and all records are the daily conditions of 33 asthmatic children in 2005. With regard to the second dataset, as mentioned previously, the adopted environmental factors were formed by air pollution data and weather data. The air pollution data were obtained from the dataset of the Environmental Protection Administration Executive Yuan, R.O.C. [25]. It contains 15 kinds of air pollutants, which are PSI (Pollutants Standards Index), PM10 (particulate matter of $10 \mu\text{m}$ or less), SO_2 , CO , O_3 , NO_2 , indicators of the quality of air pollutants, and daily concentration of pollutants reports. The weather data include daily average temperature, average relative humidity, and daily highest and lowest temperature values, which are the four important attributes and were collected by the Central Weather Bureau [24] in Tainan, Taiwan in 2005. Table 5 lists all attributes in our real datasets in detail.

According to the design of our proposed methods, we must symbolize all numerical values of the dataset to simplify the following operations.

Table 6 shows the relationship between patient respiratory flow and the risk level of asthma attacks. In the asthma dataset, there are two important indicators: (1) peak expiratory flow rate (PEFR) and (2) asthma medication dosage instructions. The PEFR is the value measured from a patient's breath

Table 5 – Attribute list of our dataset.

Sub-dataset	Category	Attributes	Count
Asthma allergic dataset	Patients' diagnostic record	ssn, d.ssn, age, reference, date	5
	Asthma symptoms	day_record, day_pefr, night_record, night_pefr, fever, night_symptom, day_symptom, nose_symptom1, nose_symptom2, nose_symptom3, nose_symptom4, eye_symptom, skin_symptom, grade, asthma_medicine_instructions, a_rhinitis, a_conjunctivitis, a_dermatitis	18
Environmental dataset	Air pollutants	Date, PSI, indicators of the quality of air pollutants, daily concentration of pollutants reports, SO ₂ , NO ₂ , O ₃ , CO, PM10, NO, NitroOxy, NHHC, HydraCarbon, HydraCarbon2, Temperature, pH	16
	Weather data	Relative humidity (%), temperature (°C), absolute maximum temperature (°C), absolute minimum temperature (°C)	4

Table 6 – PEFR value vs. its corresponding medication instructions and lights.

PEFR value (peak expiratory flow rate)	Medication instruction	Light
The value is between 80% and 100% of personal optimal value or normal estimates Variability: <20% Meaning: the state of asthma is under control	Do not need medication	Green
The value is between 60% and 80% of personal optimal value or normal estimates Variability: 20–30% Meaning: asthma symptom: the high attack, activity reduction, cough, stridor, chest tightness	To use short-acting inhaled expansion of tracheal twice each 20 min. If there is no improvement of respiratory flow rate after using 20–60 min, please go for medical treatment as soon as possible	Yellow
The value is between 60% and 80% of personal optimal value or normal estimates Variability: >30% Meaning: The state of asthma would affect patient's normal routines and activities	To use short-acting inhaled expansion of tracheal immediately. If the respiratory of flow is still lower than 60% after using, please go for medical treatment as soon as possible	Red

every morning using a peak expiratory flow meter. According to the patient's physiological conditions and the proper normal respiratory flow rate of his/her physiological conditions, patients' circumstances can be classified into the following three levels of red, yellow and green lights in the proposed mechanism. The light level and patients' allergic symptoms indicate which medication instructions should be adopted. Therefore, we symbolized a patient's physiological conditions by the values of these two indicators. Combinations of their

possible values are represented by six categorical symbols and we took them to symbolize patients' asthma statuses in our experiments of this work.

In environmental dataset, the values of its data are numeric. It is necessary to symbolize them to increase the efficiency of classification model building and the readability of prediction results. Therefore, we refer the deputy index values of each air pollutant, which are defined by proficient practitioners. The symbolized values simplify the complexity

Table 7 – Concentration of air pollutants and deputy index values.

PSI	PM10: the average value of 24 h	SO ₂ : the average value of 24 h	CO: the maximum value of the average values of 8 h	O ₃ : the maximum value in 24 h	NO ₂ : the maximum value in 24 h
	Unit: µg/m	Unit: ppb	Unit: ppm	Unit: ppb	Unit: ppb
<i>Pollutant Standards Index (PSI) and concentration of air pollutants</i>					
50	50	30	4.5	60	
100	150	140	9	120	
200	350	300	15	200	600
300	420	600	30	400	1200
400	500	800	40	500	1600
500	600	1000	50	600	2000
<i>Pollutant Standards Index (PSI) and health</i>					
PSI	0–50	51–100	101–199	200–299	≥300
Effect	Good	Moderate	Unhealthy	Very unhealthy	Hazardous

Table 8 – Level of daily extreme temperature difference table.

Extreme temperature difference (unit: °C)	Level
<2	0
>2 and <3	1
>3 and <5	2
>5 and <8	3
>8 and <13	4
>13 and <21	5
>21	6

of data and become effortlessly comprehensible. Table 7 shows a part of the deputy index values of air pollutants.

In addition to the air pollutant dataset, another dataset used in the environmental database is about the weather. The weather database contains daily information, such as extreme high and low temperatures, averaged temperature and humidity values. Each value is labeled as “up”, “down” or “flat” by referring the difference from the value of the day before. Furthermore, the difference between daily extreme high temperature and daily extreme low temperature were gathered. We call the difference value as daily extreme temperature difference and list its symbolization table in Table 8.

4.2. Experimental design

Several experiments were designed to evaluate our proposed methods and all these experiments consist of two parts. In the first part, the experiments were designed to evaluate all parameters of our methods. We discussed the impact of the length of observation interval λ on accuracy. Subsequently, we did experiments to determine a suitable value of the fair parameter of a PBCAR classifier. In the second part, all experiments were designed to verify the accuracies and performances of both methods. We firstly simulated them on the asthma related dataset, which consists of environmental data and patient allergic data. Then, we respectively examined them on different asthma related data.

In all experiments, we adopted two indicators to measure experimental results, which are the accuracy and recall. Accuracy represents the rate of the data transactions which are classified correctly, while Recall means the rate of the data transactions within one specific class, and which are classified correctly. In the asthma case, we only focused our recall evaluation on the high-risk class. Fig. 15 shows the definitions of accuracy and recall with a confusion matrix.

Confusion matrix	High-risk	Normal
Predicted as high-risk	TP	FP
Predicted as normal	TN	FN

$$\text{Accuracy} : \frac{TP+FN}{TP+TN+FP+FN}$$

$$\text{Recall} : \frac{TP}{TP+FN}$$

Fig. 15 – Definitions of accuracy and recall.**Table 9 – All parameters for our conducted experiments.**

Parameter	Symbol	Default value
Observation interval	λ	5
Balance parameter	k	400
Support of sequential pattern	$P1_support$	0.01
Support of CBA	$PBC_P2_Support$	0
Confidence of CBA	$PBC_P2_Confidence$	10
Rule number	$PBC_P2_RuleNum$	8000
Item number	$Rule_ItemNum$	8
Pruning	–	No
Using short rule	–	Yes
Training rate	$Training_Rate$	70

4.3. Evaluation of main parameters

Table 9 shows all parameters and their default values used in the experiments. The parameter λ is the observation interval for asthma attack prediction. Its default value is set as five days by doctors' suggestion and we have shown its experimental evaluation in the next paragraph. The second parameter k is the balance parameter, which is applied to reduce the impact of excessive negative class rules. We also show how we determine its default value with an experiment in the following paragraph. The third to seventh parameters are the thresholds for both methods. The support of feature extraction is the third parameter, which makes sure that the discovered features are frequent. The next two parameters are defined in the mining tool we adopted, which is named the CBA-DMII [23] and originally proposed by Liu et al. [13]. The default values of these parameters are determined by many simulated experiments in order to ensure that the discovered rules of the PBCAR are significant for asthma attack prediction. The last parameter is the training rate of a dataset. In the general data mining case, we set 70% of the whole dataset as the training constituent and 30% as the testing one. Our experimental evaluations also proved that this training rate can produce the best result on the asthma dataset.

4.3.1. Evaluation of observation interval λ

Since all related data we considered are time series datasets, the parameter observation interval λ for asthma attack prediction becomes an important factor. We discussed this end with doctors and they suggested that the observation interval λ should be set to about five days. In other words, doctors realize that all factors that occurred five days previously did not affect an asthma attack. Thus, we did this experiment to check if it is possible to make observation interval longer for collecting more features for asthma attack prevention. Fig. 16 presents the accuracy of asthma attack predictions under different observation interval length settings. The accuracy of our method grows down by increasing length of the observation interval. The reason of this trend would be that the longer observation interval makes more irrelevant information and causes the classification inaccurate. The accuracy results are over 90% when the parameter is set less than five days and the curve of the accuracy goes down when λ is over 5. Even if the experimental result with $\lambda = 3$ is the best one, it may be too short to discover complete reasons about asthma attacks. Discovering all possible features for asthma attack prevention

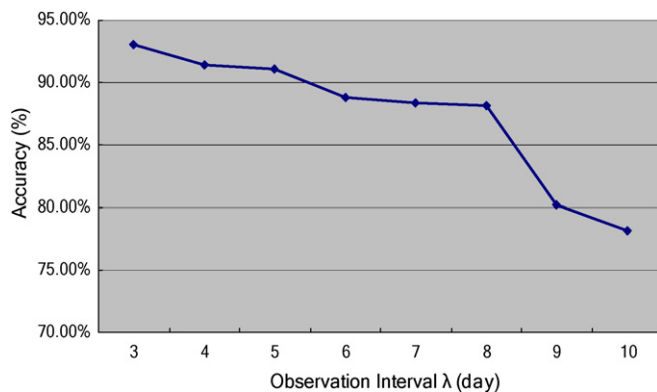


Fig. 16 – Accuracy results by varying observation interval λ .

is the purpose of our proposed mechanism. Longer λ is better and the accuracy result of $\lambda=5$ is acceptable. Consequently, we determined $\lambda=5$ as default setting for observation interval in our approach.

4.3.2. The experimental results of the balance parameter k of PBCAR

Our PBCAR method referred [14] in performing classification by scoring methods. In our applied dataset, the quantity of high-risk cases is very less than the quantity of normal cases. Therefore, the proper value of the balance parameter k of scoring method must be evaluated carefully. Fig. 17 shows the accuracy values and the recall values of varying setting of k on the asthma dataset. The recall goes up with increasing k . The accuracy increases slightly when the value of k is small, but the values of the accuracy are almost the same in the whole experiment. After the value of k is set higher than 400, both indicators do not change at all. Consequently the score of normal is not important in the asthma dataset and we can perform classification just by the score of abnormal. In summation, the features of normal segments are general features of the whole dataset, and they can also be extracted from abnormal segments. In contrast, we can predict an asthma attack when any feature of abnormal appears in one record.

4.4. Comparisons of PBDT and PBCAR

The whole asthma related dataset consists of weather, air pollutant and asthma allergic datasets. In this sub-section, we

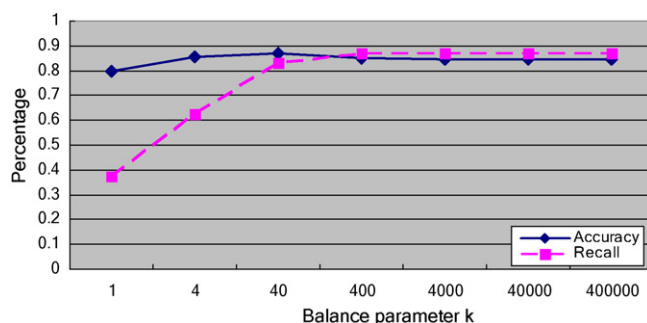


Fig. 17 – Accuracy results of our PBCAR by varying balance parameter k .

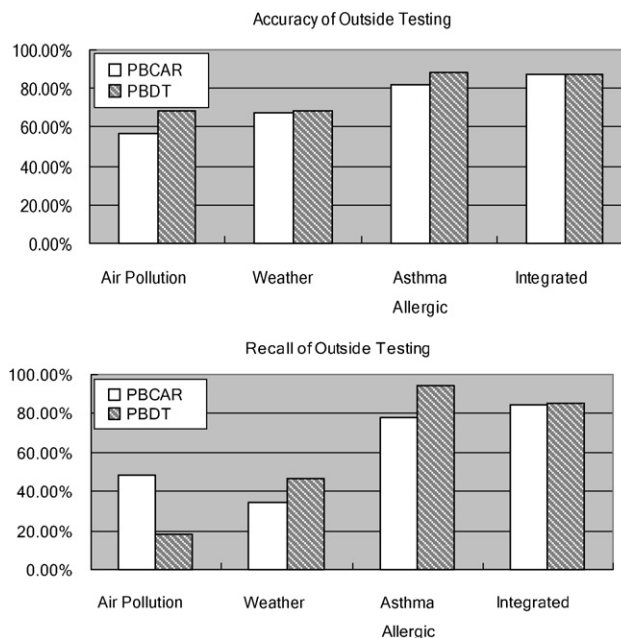


Fig. 18 – Experimental results of both proposed methods on different datasets.

evaluate the recall and accuracy of both proposed methods by applying different types of datasets and endeavor to demonstrate each factor dataset considered are useful for asthma attack prediction. In other words, we performed both methods on the weather dataset, the air pollutant dataset, the asthma allergic dataset and the asthma related dataset, which integrates all single factor datasets. Fig. 18 shows the accuracy and the recall of each dataset respectively. Overall, the results of the integrated dataset are better than the results of other single factor datasets. Although the recall of the PBDT is especially different, we inferred that the greedy architecture of the decision tree makes the result easier to be confused and unstable by inputting more information. Therefore, we discussed the two major problems in these experimental results in the following. Firstly, we compared the classifiers of both proposed methods in detail to ascertain what makes the experimental results different from ideal inference. Secondly, we discussed why the results of integrated dataset are not better than those of asthma allergic dataset using the PBDT method.

In Fig. 18, the PBCAR shows 86.89% of accuracy and 84.12% of recall, and the PBDT shows 87.52% and 85.59% of them respectively. As we can observe, the accuracy and the recall of the PBDT are slightly better than those of the PBCAR. This result is contrary to what we expected. In general, the rule-based classifier, like the CBA, can take more information into consideration than the decision tree-based classifier and perform better on doing classification. This concept is also described clearly in [22]. Tree architecture and greedy algorithm makes decision tree-based classifier lose some information. In the simulated experiments, we found that the number of rules resulted by the PBCAR is really more than the ones for the PBDT. Hence this condition may be caused by the applied dataset. Fig. 19 depicts the partial rules of the PBCAR classifier in which each rule consists of feature set and its support and confi-

• , inhaled maintenance medicine and bronchodilator → High-risk [sup: 14.16%, conf: 100%]
• , no need medicine → Normal [sup:5.584%,conf:100%]
○ , no need medicine, difference in temperature: 2°, the indicator of pollutants: O ₃ → Normal [sup: 5.076%, conf:100%]
○ ,no need medicine, • ,no need medicine, humidity: pattern (up, flat) → High risk [sup:0.761%, conf: 100%]
○ ,inhaled maintenance medicine or oral maintenance medicine day and night, ○ ,inhaled maintenance medicine and oral maintenance medicine day and night, The quality of air: good → general → High-risk [sup: 2.792%, conf: 100%]
○ ,inhaled maintenance medicine or oral maintenance medicine day and night, ○ ,inhaled maintenance medicine and oral maintenance medicine day and night, difference in temperature: 8° → High-risk [sup: 3.553%, conf: 100%]
○ ,inhaled maintenance medicine or oral maintenance medicine day and night, the quality of air: general → good → bad, PM ₁₀ :50→150 → High-risk [sup: 7.614%, conf:100%]

Fig. 19 – Partial rules of the PBC.

dence. Fig. 20 shows a classifier of the PBDT, which is in tree architecture. The tree shown in Fig. 20 is skewed meaning some features keep very high weights for classification and that additional information would cause more classification errors.

4.4.1. Analysis of trees generated by PBDT

In this section, we discussed the experimental results of the PBDT in detail and try to discover the reasons for useless environmental information. From studying the results of the PBDT presented in Fig. 18, we observe two major conditions. Firstly, the accuracy of two environmental datasets is larger than 60%, but the recalls for high-risk are very low. This situation indi-

cates both environmental datasets are not particularly related to asthma attacks but can still provide some features about asthma attack prediction. Secondly, the result of asthma allergic dataset is better than the result of the integrated one. In a generalized case, the classifier considers more related information and can perform better, but this condition shows some varying results. We checked the detail of the classifier to find reasons for this occurrence. Fig. 21 shows the classifier of the asthma allergic dataset and the classifier of the integrated one. We endeavored to compare the difference between both classifiers. There are just few differences in their sub-trees and they are located near the leaves. According to this evidence, we can infer that environmental factors contain some



Fig. 20 – Conducted rules of the PBD.

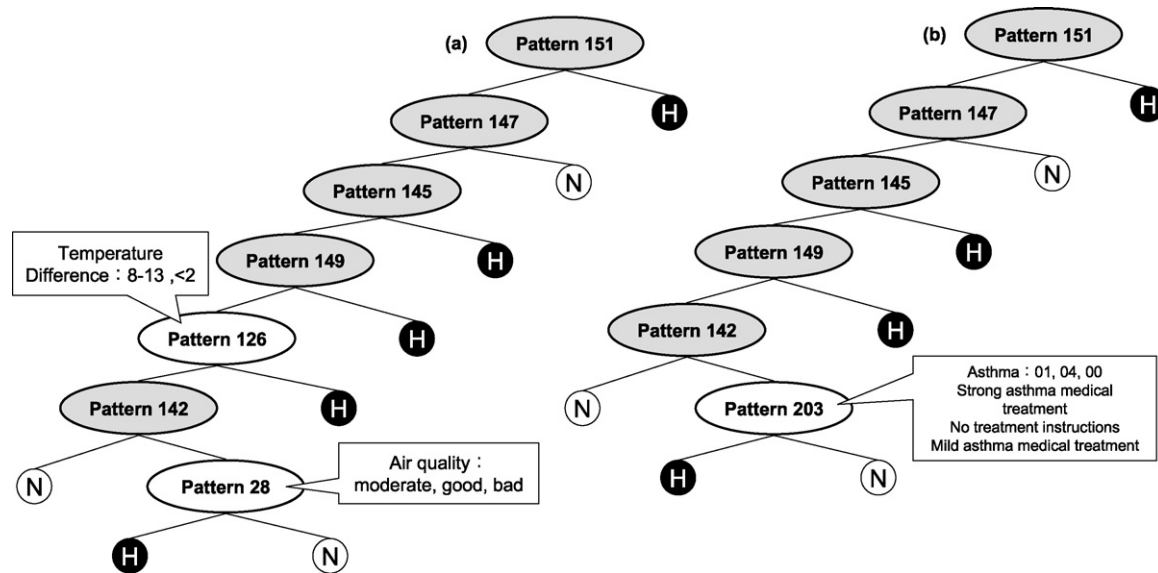


Fig. 21 – Trees of two classifiers for asthma allergic dataset and integrated dataset.

information for asthma attack prediction, but their effect is too insignificant to make great improvement in performance. With regard to recall, it is because the classifier of asthma allergic dataset consists of high-risk features. In addition, we define the recall indicator based on high-risk class. Thus, the second unexpected condition would be possible.

The integrated bio-signal mining mechanism is designed for chronic diseases. Its purpose is to integrate diverse datasets to predict attacks of a chronic disease. The experimental evaluations demonstrated our proposed design can perform well. The environmental data significantly enhanced the accuracy and the recall for asthma attack prediction. However, we adopted the daily weather and air pollutant records of a whole city as the environmental data of each patient. This kind of data may be too vague and imprecise to enhance the performance of the prediction system effectively. Sometimes, the conditions of the patient's location are very different from the environmental conditions of a whole city; for example, a building with air conditioning. These cases deterred the mechanism from effectively enhancing the performance of the asthma attack prediction system. Overall, the proposed mechanism does perform effectively and efficiently by integrating more related information.

4.5. Application for asthma attack alert

As an application, we collaborated with Tainan Asthma Allergic Children Health Association [26] and applied our proposed mechanism on the Blue Angel Asthma Home-Care System. In this system, patients record their personal daily conditions through the web service of the system. The data in the previous five days are taken by the system for analysis, and then the possibility of asthma attacks on the next day is predicted. In addition, we provided patients with helpful information for alerting them of possible future attacks, assuming that they are likely to suffer from an asthma attack. According to doctors' observations in real implementations, many patients felt

that our proposed mechanism is really helpful and the risk prediction can effectively help them to reduce the incidence of dangers.

5. Conclusions and future works

In this paper, we proposed an integrated bio-signals mining mechanism, which considers patients' clinical data and the environmental information where patients are domiciled. With respect to implementation, two methods named PBDDT and PBCAR were proposed. Their mechanisms consist of two phases, which make processing these different datasets together possible. The PBDDT integrates the concepts of sequential pattern mining and decision tree mining to build a decision tree with all asthma related features for asthma attack prediction. The tree architecture makes the prediction rules easier to understand for general users. In addition, the PBCAR merges the concepts of sequential pattern mining and classifiable association rule mining. It elicits the association rules between the prefix features of all asthma attacks of the integrated dataset. The PBCAR predicts risks with these association rules. The corresponding rules are more detailed and are suitable for medical healthcare practitioners for in-depth studies of chronic diseases. Through experimental evaluations, these two methods were proven to work effectively. The proposed mechanism evaluated the concentration of air pollutants, climatic changes in the weather, and patient's asthma related allergies to provide an asthma attack prediction and some medication instructions instantaneously. In conclusion, these outputs help patients and doctors prevent asthma attacks effectively and efficiently. There are several novel characteristics in our proposed mechanism and can be summarized as the following:

1. Method for integrating multi-format information: We proposed the PBDDT and PBCAR methods, which integrate the concepts of traditional mining techniques. The designs

make the implementation more intuitive and transform our mechanism appropriate to be applied on the dataset with a complex format. Simultaneously, the results of our proposed mechanism are easier to apply and to understand for users.

2. Consideration of compound factors in patient monitoring: We not only discussed patients' asthma symptoms, but also took air pollutant database and weather dataset into consideration. More useful information makes the proposed mechanism more comprehensive.
3. Accurate prediction mechanism: Through evaluation of all related data, our proposed methods can predict the probability of asthma attacks in near future. Experimental results proved that our prediction mechanism has high accurate. This mechanism evaluates patients' daily health statuses from their daily records and effectively alert patients for possible asthma attacks in near future.
4. Provision of readable reference information: Our mechanism integrates several benefits of data mining methods. Thus, both methods can discover readable features and build classification rules. These classification rules are easy to read and able to offer patients and healthcare workers as a reference of patients' illnesses.

For future work, we propose to explore and to apply the proposed mechanism to other kinds of chronic diseases. In addition, we will take more chronic disease related factors into consideration, such as periodic weather changes, the relationship between the features of different factor datasets. Furthermore, we will take into account personal physiological profiles such as DNA data and eating habits of patients, with the aim of providing personalized preventive medical care.

Conflict of interest

None.

Acknowledgement

This research was supported by the Landmark Project of National Cheng Kung University, Taiwan, ROC.

REFERENCES

- [1] R. Agrawal, R. Srikant, Fast algorithms for mining association rules, in: The 20th International Conference on Very Large Databases, Santiago, Chile, September, 1994.
- [2] R. Agrawal, R. Srikant, Mining sequential patterns, in: Eleventh International Conference on Data Engineering, 1995.
- [3] M. Barua, H. Nazeran, P. Nava, B. Diong, M. Goldman, Classification of Impulse Oscillometric Patterns of Lung Function in Asthmatic Children Using Artificial Neural Networks, Engineering in Medicine and Biology Society, 2005.
- [4] B.V. Bonnländer, A.S. Weigend, Selecting input variables using mutual information and nonparametric density estimation, in: Proceedings of the International Symposium on Artificial Neural Networks (ISANN'94), 1994.
- [5] T.H. Cheng, C.P. Wei, V.S. Tseng, Feature selection for medical data mining: comparisons of expert judgment and automatic approaches, in: Proc. 19th IEEE Int'l Conf. on Computer-based Medical Systems (CBMS'06), USA, June, 2006.
- [6] H.T. Chu, C.C. Huang, Z.H. Lian, J.J.P. Tsai, A ubiquitous warning system for asthma-inducement, in: IEEE Sensor Networks, Ubiquitous, and Trustworthy Computing, 2006.
- [7] O. Cure, Designing patient-oriented systems with semantic web technologies, in: IEEE Symposium on Computer-Based Medical Systems (CBMS), 2003, p. 195.
- [8] J. Finkelstein, M.R. Cabrera, G. Hripcsak, Web-based monitoring of asthma severity: a new approach to ambulatory management, Proceedings of IEEE International Conference on Information Technology Applications in Biomedicine (1998) 139–143.
- [9] I. Guyon, A. Elisseeff, An introduction to variable and feature selection, The Journal of Machine Learning Research Archive 3 (2003) 1157–1182.
- [10] R.L. Jan, J.Y. Wang, M.C. Huang, S.M. Tseng, H.J. Su, L.F. Liu, An Internet-based interactive telemonitoring system for improving childhood asthma outcomes in Taiwan, Telemedicine and e-Health (2006) 1–28.
- [11] K. Kira, L.A. Rendell, A practical approach to feature selection, in: Proc. 9th international workshop on Machine Learning, 1992, pp. 249–256.
- [12] N. Lesh, M.J. Zaki, M. Ogihara, Mining features for sequence classification, in: Proc. 5th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining, San Diego, CA, USA, 1999, pp. 342–346.
- [13] B. Liu, W. Hsu, Y. Ma, Integrating classification and association rule mining, in: The Fourth International Conference on Knowledge Discovery and Data Mining, New York, USA, 1998.
- [14] B. Liu, Y. Ma, C.K. Wong, P.S. Yu, Scoring the data using association rules, Applied Intelligence 18 (2) (2003) 119–135.
- [15] M. O'Leary, D. Koolpiruck, W. Balachandran, J. Emberlin, R. Lewis, The role of electrostatic charge accumulated by respirable sized allergens with regard to thunderstorm asthma, in: Industry Applications Conference, 2005.
- [16] J.R. Quinlan, Induction of Decision Trees, Machine Learning 1 (1) (2003) 81–106 (Kluwer Academic Publishers).
- [17] S. Ruggieri, Efficient C4.5, IEEE Transactions on Knowledge and Data Engineering 14 (2) (2002) 438–444.
- [18] C.L. Stokes, A.K. Lewis, T. Paterson, C.C. Leong, N. Defranoux, G. Bub Entelos, Asthma Physiolab: a dynamic, computer-based mathematical model of acute and chronic asthma, in: The First Joint BMES/EMBS Conference Serving Humanity, Advancing Technology, Akmta, GA, USA, 1999.
- [19] V.S. Tseng, L.C. Chen, C.H. Lee, J.-S. Wu, Y.-C. Hsu, Development of a vital sign data mining system for chronic patient monitoring, in: Proc. 2008 International Workshop on Intelligent Informatics in Biology and Medicine (IIBM-08), Spain, March, 2008.
- [20] V.S. Tseng, C.H. Lee, J.C. Chen, An integrated data mining system for patient monitoring with applications on asthma care, in: Proc. of the 21th IEEE International Symposium on Computer-Based Medical Systems (CBMS 2008), Finland, June 17–19, 2008.
- [21] P. Varady, Z. Benyo, B. Benyo, An open architecture patient monitoring system using standard technologies, IEEE Transactions on Information Technology in Biomedicine 6 (1) (2002).
- [22] A. Veloso, W. Meira Jr., M.J. Zaki, Lazy associative classification, in: The Sixth International Conference on Data Mining, 2006, pp. 645–654.

-
- [23] CBA-DMII, Developed by National University of Singapore. [<http://www.comp.nus.edu.sg/~dm2/index.html>], 2005.
- [24] Central Weather Bureau, R.O.C. [<http://www.cwb.gov.tw/>], 2005.
- [25] Environmental Protection Administration Executive Yuan, R.O.C. [<http://edb.epa.gov.tw/>], 2005.
- [26] Tainan Asthma Allergic Children Health Association, R.O.C. [<http://140.116.58.191/asthma/index.php>], 2005.