ARB: Knowledge Discovery and Disease Diagnosis on Thyroid Disease Diagnosis integrating Association Rule with Bagging Algorithm

Dongyang Li, Dan Yang, Jing Zhang

Abstract—**Mastering disease influence factors promises to advance clinical research and provides a possible decision making. In this paper, we propose a framework ARB, which is integrating association rule mining algorithm with bagging algorithm. ARB consists of two main modules 1) knowledge discovery and 2) disease diagnosis. Firstly association rule mining algorithm is used to investigate the sick and healthy factors which contribute to disease for males and females. This also aims to select the most robust and effective features to reduce the dimensions. And then we use ensemble algorithm to diagnose disease based on the data filtered by the first module. The framework ARB applies three real thyroid datasets in UCI machine learning repository. Though the association rules generated by Apriori algorithm, we know thyroid disease have different effects on people of different age intervals, and the elderly from 60 to 80 are the most likely to suffer from thyroid disease. The results also show that the two age intervals (30, 40] and (50, 60] are the age intervals with the highest recurrence rate of thyroid disease. And for gender factor, men have more chances of being free from thyroid disease than women. For women in their twenties, they have less risk. After that, we use thyroid disease knowledge from these rules as the input of model for diagnosing thyroid disease. The experimental results significantly show that the performance of ARB outperforms others, which also shows the feasibility and practical value of the framework ARB in thyroid aided diagnosis.**

Index Terms*—*Thyroid disease; Association rule mining; Apriori algorithm; bagging algorithm

I. Introduction

Disease diagnosis is coming to a new era where abundant diagnosis data are applied to obtain efficient features and building the effective diagnosis model. And the Endocrine Branch of the Chinese Medical Association has announced that thyroid disease has become the second largest disease in endocrine disease besides diabetes mellitus [1], and about 30% of young (0-44) and middle-aged people (45-59) and more than 50% of the elderly (60-90) [2] will be associated with it each year [3]. According to the latest research by Taylor et al., the main causes of thyroid disease may be: gender, insufficient iodine intake, excessive iodine intake, the transition from iodine deficiency to adequate iodine intake, other autoimmune conditions, genetic risk factors, smoking, alcohol consumption, drug abuse, selenium deficiency, infection and syndrome [4]. And study [5-8] also show that sex, age and weight have an impact on thyroid function. Therefore, exploring the association rules among those features is a fundamental task which can develop the medical diagnosis.

In recent years, with the advent of the era of big data, artificial intelligence algorithms such as machine learning and artificial neural networks have made outstanding contributions in various fields. Compared with machine learning algorithm, the construction of neural network model is relatively complex, and it operates in the 'black box'. The connection weights between neurons are meaningless, and the interpretability is not high. Unlike other fields, the diagnosis indicators in the medical field contain important hidden information. Association rule mining is a data mining technology, which is mainly used to discover the relationship among different attributes. And WHO has found that data mining algorithms can greatly improve some problems in the medical field [9].

II. Related work

As data mining technology become more and more mature, we will review many main literatures. Samo Riyanarto et al. [10] used generated positive and negative rules applied for compliance checking towards the test dataset. Ogunde et al. [11] used association rule mining algorithms to set a system to adapt to constantly changing databases and mining environments. Kaoungku Nuntawut et al. [12] used association rule mining algorithms to select features. Wang Yingquan et al. [13] applied association rule mining algorithms in business. In medical field, Zhang Yang et al. [14] analyzed the distribution rules and relevant relation of TCM signs, symptoms and syndrome elements in essential hypertension. They collected the signs and symptom of EH patients from the Longhua Hospital affiliated to Shanghai University of Traditional Chinese Medicine, Putuo Hospital affiliated to Shanghai University of Traditional Chinese Medicine and Shanghai Wanggang

Dongyang Li, is with School of Computer Science and Software Engineering, University of Science and Technology LiaoNing, Anshan, China (e-mail: hclidonyang@ 163.com)

Dan Yang, the corresponding author, is associate professor with School of Computer Science and Software Engineering, University of Science and Technology LiaoNing, Anshan, China (e-mail: asyangdan@ 163.com)

Jing Zhang, is with School of Computer Science and Software Engineering, University of Science and Technology LiaoNing, Anshan, China (e-mail: zwinerj@ 163.com)

Community Hospital from April 2014 to October 2015. They used SPSS Statistic to analyze data and Apriori algorithm was used to extract the association rules. Conclude that there were certain regularities in the distribution of TCM SEs and SSs in EH patients. Some SEs and SSs had the core role in the diagnosis of EH. In related work [15], a model was established of Xinan Wangs internal medicine for treating epigastric pain through the data mining technique, and provided a more sufficient scientific basis for systematically discussing the rule of traditional Chinese medicines in treating epigastric pain. SPSS Statistic was also used to analyze data, and Apriori algorithm was applied to extracting the core prescription. The experimental results showed that Wangs doctors in Xinan often treated the stomach and spleen with the same treatment of liver and spleen. Wangs physician treated epigastric pain with both liver and spleen. Wangs physician assisted the spleen with Tongyang method, or relieved phlegm, dredged liver and Qi to Tongyang, or Xinwen Sanhan to Tongyang, or Jianpi Huashi to Tongyang, or Huoxue Xingqi to Tongyang and Huoxue Sanjie method and Huoxue Huayu method. Paper [16] proposed a classification model based on atomic classification association rules, and applied it to construct the classification model of a Tibetan medical syndrome for the common plateau disease called Chronic Atrophic Gastritis. They used the constraint-based Apriori algorithm to mine the strong atomic classification association rules between symptoms and syndrome. Then they established the classification model of the Tibetan medical syndrome, and the idea of partial classification to predict this Tibetan medical syndrome. The experimental result showed that the accuracy of the model always had a better performance. In related work [17], Dong Wenzhe et al. were performed to evaluate the effects of different dosage forms of Tripterygium wilfordii on immune inflammatory metabolic markers in patients with rheumatoid arthritis. Apriori algorithm was used to analyze the medical records of hospitalized patients. In related work [18], Apriori algorithm was used to extract rules of DING Gan-ren in the following aspects: common syndromes, common external diseases, some spectacular herbs selection of certain symptoms and some distinctive herb-pairs.

Bagging ensemble algorithm is also used in various fields. Lv Yanxia et al. [19] used bagging ensemble algorithm for big data stream learning. In related work [20], Liu Hailing and Zang Xian proposed a pattern recognition algorithm of automatic identification of epithelial cell which used bagging ensemble algorithm. In the field of medical big data, Dai Peng et al. [21] presented an automatic diagnosis which used bagging ensemble algorithm for diagnosis and prognostication of Alzheimer's disease. Experimental results showed that the proposed algorithm yielded superior results compared to the other methods, suggesting promising robustness for possible clinical applications. In related work [22], Prata Marco et al. predicted the Medical Specialty (MS) discharge in a hospital Urgency Department (UD) by bagging ensemble algorithm. The experimental results were achieved using a REP-Tree base algorithm and a ten-fold cross-validation, achieving 91.96 % of accuracy and 0.85 of F1-score.

III. METHODOLOGY

A. *ARB overview*

Traditional Chinese medicine has been developed over thousands of years, and it is very complex. Over these years, traditional Chinese doctors have summarized different factors which contribute to disease. It is time-consuming, and this meaningful information needs further proof. Therefore, we promote a knowledge discovery and disease diagnosis framework ARB which integrated association rule mining algorithm and bagging algorithm. The framework simulates the diagnosis process of traditional Chinese medicine, and the framework is shown in Fig.1.

Most feature selection methods, including chi-square test, relief algorithm, logistic regression, and so on, select features by setting thresholds manually. However, in the medical field, the relations between disease and diagnosis indicators are very important for the correctly disease diagnosis. So association rule mining algorithm, i.e. Apriori algorithm, is used to mine meaningful disease knowledge, and we summarize this knowledge and select most frequent attributes as ARB attributes. Then bagging algorithm is used to diagnose disease. In this way, we can not only select attributes, but also analyze the relations between disease and diagnosis indicators. And the framework ARB increases the accuracy of diagnosis actually.

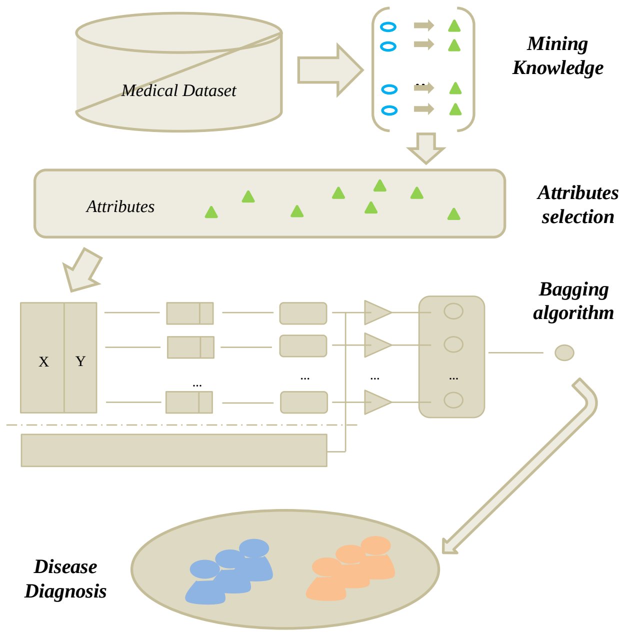

Fig.1.The framework of ARB

B. *Classification of thyroid disease*

In general, there are two classification standards for thyroid disease. In first classification standard, thyroid disease can be divided into medical treatment of thyroid disease. Another can be divided into surgical treatment of thyroid disease. Furthermore, medical treatments of thyroid disease mainly include hyperthyroidism, thyroiditis, hypothyroidism, and so on [23]. Surgical treatments of thyroid disease mainly include thyroid cyst and thyroid tumor, which seriously threaten health. In this paper, we aims at extracting association rules treated by medical treatment, which uses thyroid disease dataset, and we use a uniform name below as thyroid disease.

C. *Apriori algorithm*

Association rule mining is a recognized data mining technology [24]. The form of the rule generated by the association learning is "LHS (left-hand-side) ⇒ RHS (right-hand-side)", where LHS and RHS are disjoint itemsets. This rule indicates that the RHS itemset is likely to occur whenever the LHS item set occurs. Support and Confidence are two indicators that measure the rules, which reflect the validity and certainty of the rules [25].

For the database transactions $D = \{I_1, I_2, I_3, \cdots, I_m\}$, let X, Y be an item respectively, and $A \subset D$, $B \subset D$, $A \neq \emptyset$, $B \neq \emptyset$, $A \geq B$. Then Rule '$A \Rightarrow B$' is established in D, and Support and Confidence are as follows (formula 1 and formula 2):

$$Support\ (X \leq Y) = P(X \cap Y) \quad (1)$$

$$Confidence(X \leq Y) = P\ (Y|X) = P(X \cap Y)/P(X) \quad (2)$$

Many association rule mining algorithms have been proposed, and the generated rules are different by different algorithms [26]. Apriori algorithm is one of the most influential mining algorithms for frequent itemsets [27]. In Apriori algorithm, the validity and authenticity of mining results largely depend on the choice of minimum support. Setting the minimum support too high or too low will affect the generated rules, and it is difficult to obtain satisfactory results without sufficient application experience. According to literature [28], we set the upper limit of minimum support to 0.2 in this paper, the lower limit of minimum support is set to 0.1, and the minimum confidence is set to 0.95. And pseudocode is shown in Table I.

D. *Bagging algorithm*

Due to the unbalanced characteristic of medical data, the total number of healthy individuals is obviously more than sick individuals, and the accuracy of predictive diagnosis of single algorithm is very limited in medical diagnosis. So ensemble algorithm becomes one of the methods to solve this problem. The ensemble algorithm is to complete the learning task by building multiple base algorithms, which can be traditional machine learning classification algorithms such as KNN algorithm, or artificial neural network algorithm. In ensemble algorithm, bagging algorithm has obvious advantages in reducing over fitting, so it usually performs well in strong algorithm and complex model. The process of bagging algorithm is as follows: first of all, M samples are randomly selected from the original data set D and repeated T times, then T training sets are generated. Each training set can train a base algorithm, and finally T algorithms are generated. The prediction results will be determined by these algorithms (the most categories in the voting results of the algorithm are selected as the final prediction results), the flow chart of bagging algorithm is shown in Fig.2, and the pseudocode is shown in Table II.

TABLE I
THE PSEUDOCODE OF APRIORI ALGORITHM

```
Input：D: Transaction Database
       min_sup: Minimum Support Counting Threshold
Procedure：
    Procedure apriori (D, min_sup)
    // find most frequent 1- itemsets
        L1 = find_frequent_1- itemsets (D);
        For (k=2; Lk-1! = null; k++)
            Return L= All Frequent Sets;
    Step 1: Join
        Procedure Apriori_gen (Lk-1: frequent (k-1) - itemsets)
            Return Ck;
    Step 2: Prune
        Procedure has_infrequent_sub (c: candidate k-itemset; Lk-1:
        frequent (k-1)-itemsets)
            Return FALSE;
```

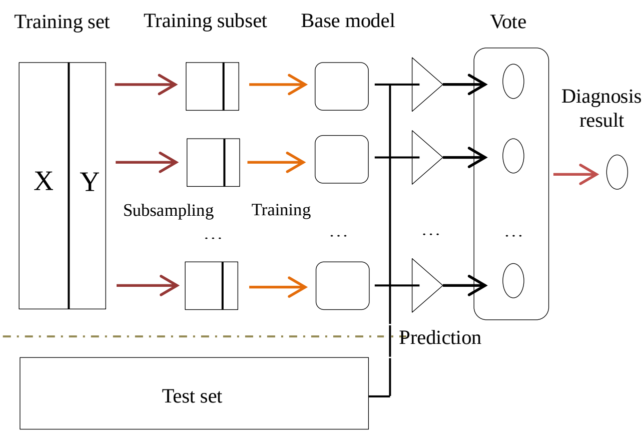

Fig.2.The flow chart of bagging algorithm

TABLE II
THE PSEUDOCODE OF BAGGING ALGORITHM

Input：Training set: $D = \{(x_1, y_1), (x_2, y_2), (x_3, y_3)\ldots(x_n, y_n)\}$;

Base algorithm：ζ

Training times：T

Classification attribute：Y

Classification result：O

Procedure：

Step 1:

Given a size of N training set D

Step 2:

Bagging: obtain T new training set D_i by sampling from D with replace, and each D_i is m in size.

Step 3:

Obtain T training sets $D_i = \{(x_1, y_1), (x_2, y_2)\ldots(x_m, y_m)\}$, use them to train the classifier, and T results are obtained.

for $t = 1, 2, 3, \ldots, T$ {

$O_t = \zeta(D_i)$ }

Step 4:

T results were voted on, with the majority of votes being final classification values.

$\mathrm{O} = arg_{\ y \in Y}\ max \sum_{t=1}^{T}(O_t\ (x) = y)$

IV. EXPERIMENTS

A. Dataset overview

The sensitive personal information are involved in disease dataset and the most important thing in research is data-related privacy issue, which is also the core issue of data sharing in the era of healthcare and big data [29]. Therefore, we use three real thyroid disease datasets in the UCI machine learning repository, which are widely used by data mining researchers.

i. *Dataset preprocessing and feature selection*

In order to compare with the accuracy of diagnosis studied by other researchers in literature [30-32], we choose the same dataset, i.e. new-thyroid dataset, in UCI machine learning repository. However, in modern medical data, there are few diagnostic indexes which only contain numerical attributes such as *TSH*, and in order to provide relatively larger amount of data, sick dataset and sick-euthyroid dataset in UCI machine learning repository are merged and processed. Then the merged dataset is used for our experiment, and we named the dataset as thyroid-data. Sick dataset consists of 29 attributes, and sick-euthyroid dataset has 25 attributes, so we choose 25 same attributes to compose thyroid-data dataset. Also due to many default values in thyroid-data dataset, the data is cleaned with SPSS first. Then SPSS is used to preprocess the age attribute, the numeric attribute is converted to the nominal attribute with an age range of 10. The details of 13 baseline attributes are shown in Table III.

TABLE III
BASELINE ATTRIBUTES OF THYROID-DATA DATASET

Attribute ID	Attribute Name	Attribute Explanation
1	*age_group*	Age interval
2	*sex*	M=Male or F=female
3	*on_thyroxine*	Whether taking thyroxine drugs, T=true or F=False
4	*query_on_thyroxine*	Whether taken thyroxine drugs T=true or F=False
5	*on_antithyroid_medication*	Whether taking anti-thyroid drugs T=true or F=False
6	*sick*	Whether sick T=true or F=False
7	*pregnant*	Whether pregnant T=true or F=False
8	*thyroid_surgery*	Whether in thyroid surgery T=true or F=False
9	*query_hypothyroid*	Whether had hypothyroidism T=true or F=False
10	*query_hyperthyroid*	Whether had hyperthyroidism T=true or F=False
11	*lithium*	Whether taking drugs containing lithium T=true or F=False
12	*goitre*	Whether have thyroid goitre T=true or F=False
13	*tumor*	Whether have thyroid tumor T=true or F=False
14	*class*	Health or sick

ii. *Application of Apriori algorithms in thyroid-data Dataset*

In this subsection, thyroid-data dataset is used to generate rules by Apriori algorithm, and all patients' individuals are divided into two categories, one is healthy classification, the other is sick classification. This subsection selects Top-10 optimal rules with a confidence higher than 95%. Firstly we set the RHS to healthy and sick classification. Then we mine the association rules based on gender. The following subsections provide more details.

B. Rules extraction through Apriori algorithm mining

In the first experiment, the generated rules by Apriori algorithm are shown in Table IV. Then we extract rules which contain sick classification in the right-hand side (RHS) independently. In order to have a better visualization analysis based on these generated rules, we integrate knowledge graph with decision tree to form a new rule description diagram. We name it as rule analysis diagram, and rule analysis diagrams which rules contain healthy rules and sick rules are in Fig.3 and Fig.4 respectively. In the rule analysis diagram defined in this section, the top rectangle represents whether it is in the state of thyroid disease or not, which is named as the state node. Ellipses are regular nodes. In each rule analysis diagram, there is only one rule in each rule node. The connection between two rule nodes is represented by a line segment with an arrow, and all the arrows point up to the node. The number near the line segment indicates times the corresponding rule appears in the mined rules. At the same time, the relationship between the state node and the rule node decreases with the increase of the distance from the state node. There are also unconnected nodes, which indicate that the attributes do not appear in the mining rules. The free curve representation in the diagram connects two nodes with another node, which is used for a clearer representation of rule analysis diagram.

From Table IV, Fig.3 and Fig.4, Top-10 association rules mined for healthy classification are all related to the age interval (30,40], which indicates that people of 30 to 40 have more chances of being free from thyroid disease. At the same time, there are also many factors than can impact them. Firstly it is an important indicator for healthy people that people have no history of hypothyroidism. And then, have no surgical treatments of thyroid disease also is a good indicator. The results also indicate that taking drugs containing thyroxine may also have an impact on health.

Considering the sick classification, all the rules are attributed to the age interval (60, 70] and (70, 80]. And eight of the ten rules generated by thyroid-data dataset indicate that women have more chances of thyroid disease.

TABLE IV
TOP-10 RULES GENERATED BY APRIORI ALGORITHM

Rules class		Rules	Confidence
Health	1.	age_group=(30, 40] ∩ sick=F ∩ thyroid_surgery=F ∩ query_hypothyroid=F ∩ query_hyperthyroid=F ∩ tumor=F ==> class=health	0.98
	2.	age_group=(30, 40] ∩ query_on_thyroxine=F ∩ sick=F ∩ thyroid_surgery=F ∩ query_hypothyroid=F ∩ query_hyperthyroid=F ∩ tumor=F ==> class=health	0.98
	3.	age_group=(30, 40] ∩ sick=F ∩ thyroid_surgery=F ∩ query_hypothyroid=F ∩ query_hyperthyroid=F ∩ lithium=F ∩ tumor=F ==> class=health	0.98
	4.	age_group=(30, 40] ∩ sick=F ∩ thyroid_surgery=F ∩ query_hypothyroid=F ∩ query_hyperthyroid=F ∩ goitre=F ∩ tumor=F ==> class=health	0.98
	5.	age_group=(30, 40] ∩ on_antithyroid_medication=F ∩ sick=F ∩ thyroid_surgery=F ∩ query_hypothyroid=F ∩ query_hyperthyroid=F ∩ tumor=F ==> class=health	0.98
	6.	age_group=(30, 40] ∩ query_on_thyroxine=F ∩ sick=F ∩ thyroid_surgery=F ∩ query_hypothyroid=F ∩ query_hyperthyroid=F ∩ lithium=F ∩ tumor=F ==> class=health	0.98
	7.	age_group=(30, 40] ∩ query_on_thyroxine=F ∩ sick=F ∩ thyroid_surgery=F ∩ query_hypothyroid=F ∩ query_hyperthyroid=F ∩ goitre=F ∩ tumor=F ==> class=health	0.98
	8.	age_group=(30, 40] ∩ sick=F ∩ query_hypothyroid=F ∩ query_hyperthyroid=F ∩ tumor=F ==> class=health	0.98
	9.	age_group=(30, 40] ∩ sick=F ∩ thyroid_surgery=F ∩ query_hyperthyroid=F ∩ tumor=F ==> class=health	0.98
	10.	age_group=(30, 40] ∩ query_on_thyroxine=F ∩ sick=F ∩ query_hypothyroid=F ∩ query_hyperthyroid=F ∩ tumor=F ==> class=health	0.98
Sick	1.	age_group=(70, 80] ∩ sex=F ∩ pregnant=F ==> class=sick	1
	2.	age_group=(70, 80] ∩ sex=F ∩ thyroid_surgery=F ==> class=sick	1
	3.	age_group=(60, 70] ∩ on_thyroxine=F ∩ pregnant=F ==> class=sick	1
	4.	age_group=(60, 70] ∩ on_thyroxine=F ∩ lithium=F ==> class=sick	1
	5.	age_group=(60, 70] ∩ pregnant=F ∩ query_hypothyroid=F ==> class=sick	1
	6.	age_group=(60, 70] ∩ query_hypothyroid=F ∩ lithium=F ==> class=sick	1
	7.	age_group=(70, 80] ∩ sex=F ∩ query_on_thyroxine=F ∩ lithium=F ==> class=sick	1
	8.	age_group=(70, 80] ∩ sex=F ∩ on_antithyroid_medication=F ∩ pregnant=F ==> class=sick	1
	9.	age_group=(70, 80] ∩ sex=F ∩ on_antithyroid_medication=F ∩ thyroid_surgery=F ==> class=sick	1
	10.	age_group=(70, 80] ∩ sex=F ∩ pregnant=F ∩ goitre=F ==> class=sick	1

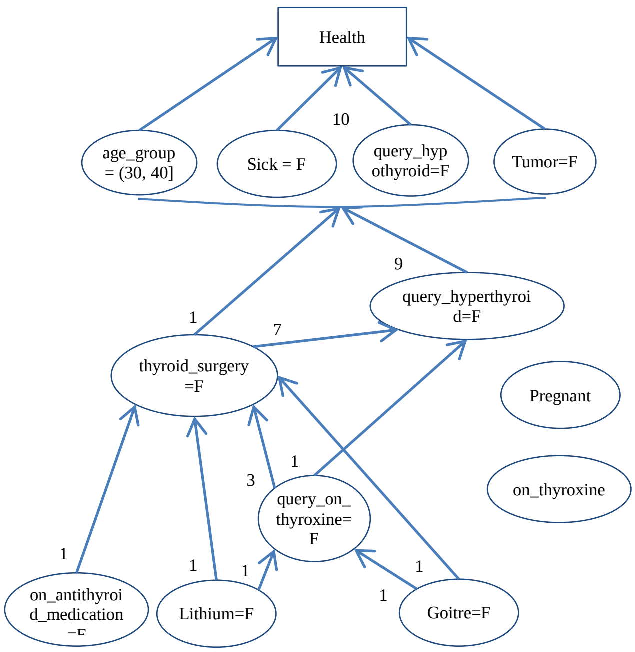

Fig.3. Rule analysis diagram for healthy rules

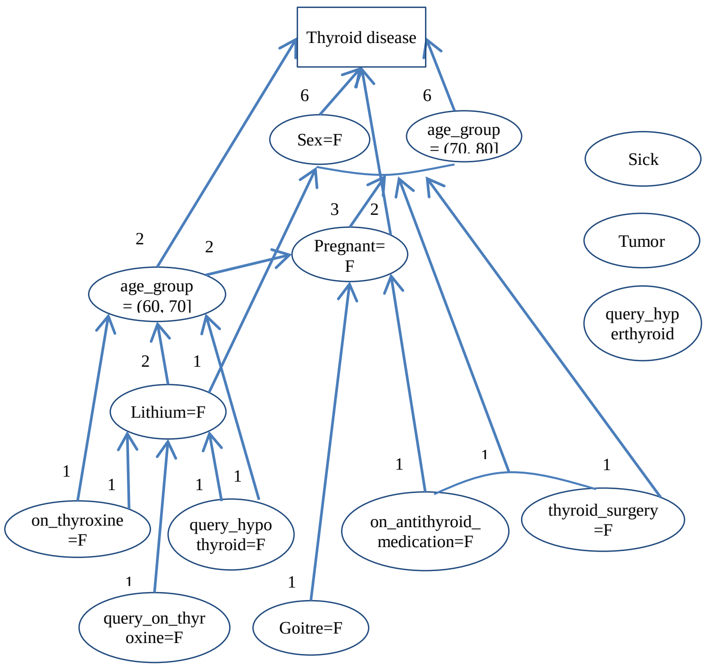

Fig.4. Rule analysis diagram for sick rules

C. *Apriori algorithm mining to detect gender conditions*

In the previous subsection, rules generated to both gender are not very specific. In this subsection, we will split the dataset according to the factors of male and female, and the rules mined for sick classification and healthy classification will be extracted again. The effect of gender in thyroid disease will be studied in more details. In this section, the factors of gender and pregnant are removed in male group and the factor of gender is removed in female group. The aim is to separately observe which factors are significantly related to thyroid disease in men and women. The rule extractions are shown in Table V and Table VI.

For men, it is confirmed again that the elderly are most likely to suffer from thyroid disease. The middle-aged men (50, 60] who has history of thyroid disease need to pay attention to the recurrence of thyroid disease, and men who are suffering from surgical treatments of thyroid disease also need to pay attention to thyroid disease. For women, people of 20 to 30 have the greatest chances of being free from thyroid disease, and 'on_thyroxine=T' is almost the only indicator to impact the health.

TABLE V
TOP-10 RULES GENERATED FOR MALE

Rules class		Rules	Confidence
Health	1.	age_group=(50, 60] ∩ thyroid_surgery=F ∩ query_hypothyroid=F ∩ query_hyperthyroid=F ==> class=health	0.95
	2.	age_group=(50, 60]∩thyroid_surgery=F ∩ query_hypothyroid=F ∩ query_hyperthyroid=F ∩ tumor=F ==> class=health	0.95
	3.	age_group=(50, 60] ∩ query_on_thyroxine=F ∩ thyroid_surgery=F ∩ query_hypothyroid=F ∩ query_hyperthyroid=F ==> class=health	0.95
	4.	age_group=(50, 60] ∩ thyroid_surgery=F ∩ query_hypothyroid=F ∩ query_hyperthyroid=F ∩ goitre=F ==> class=health	0.95
	5.	age_group=(50, 60] ∩ query_on_thyroxine=F ∩ thyroid_surgery=F ∩ query_hypothyroid=F ∩ query_hyperthyroid=F ∩ tumor=F ==> class=health	0.95
	6.	age_group=(50, 60] ∩ thyroid_surgery=F ∩ query_hypothyroid=F ∩ query_hyperthyroid=F ∩ goitre=F ∩ tumor=F ==> class=health	0.95
	7.	age_group=(50, 60] ∩ on_antithyroid_medication=F ∩ thyroid_surgery=F ∩ query_hypothyroid=F ∩ query_hyperthyroid=F ==> class=health	0.95
	8.	age_group=(50, 60] ∩ thyroid_surgery=F ∩ query_hypothyroid=F ∩ query_hyperthyroid=F ∩ lithium=F ==> class=health	0.95
	9.	age_group=(50, 60] ∩ query_on_thyroxine=F ∩ thyroid_surgery=F ∩ query_hypothyroid=F ∩ query_hyperthyroid=F ∩ goitre=F ==> class=health	0.95
	10.	age_group=(50, 60] ∩ on_antithyroid_medication=F ∩ thyroid_surgery=F ∩ query_hypothyroid=F ∩ query_hyperthyroid=F ∩ tumor=F ==> class=health	0.95
Sick	1.	age_group=(60, 70] ==> class=sick	1
	2.	age_group=(60, 70] ∩ on_antithyroid_medication=F ==> class=sick	1
	3.	age_group=(60, 70] ∩ thyroid_surgery=F ==> class=sick	1
	4.	age_group=(60, 70] ∩ goitre=F ==> class=sick	1
	5.	age_group=(60, 70] ∩ on_antithyroid_medication=F ∩ thyroid_surgery=F ==> class=sick	1
	6.	age_group=(60, 70] ∩ on_antithyroid_medication=F ∩ goitre=F ==> class=sick	1
	7.	age_group=(60, 70] ∩ thyroid_surgery=F ∩ goitre=F ==> class=sick	1
	8.	age_group=(60, 70] ∩ on_antithyroid_medication=F ∩ thyroid_surgery=F ∩ goitre=F ==> class=sick	1
	9.	age_group=(60, 70] ∩ query_on_thyroxine=F ==> class=sick	1
	10.	age_group=(60, 70] ∩ lithium=F ==> class=sick	1

TABLE VI
TOP-10 RULES GENERATED FOR FEMALE

Rules class		Rules	Confidence
Health	1.	age_group=(20, 30] ∩ on_thyroxine=F ==> class=health	0.99
	2.	age_group=(20, 30] ∩ on_thyroxine=F ∩ goitre=F ==> class=health	0.99
	3.	age_group=(20, 30] ∩ on_thyroxine=F ∩ query_on_thyroxine=F ==> class=health	0.99
	4.	age_group=(20, 30] ∩ on_thyroxine=F ∩ sick=F ==> class=health	0.99
	5.	age_group=(20, 30] ∩ on_thyroxine=F ∩ query_on_thyroxine=F ∩ goitre=F ==> class=health	0.99
	6.	age_group=(20, 30] ∩ on_thyroxine=F ∩ sick=F ∩ goitre=F ==> class=health	0.99
	7.	age_group=(20, 30] ∩ on_thyroxine=F ∩ lithium=F ==> class=health	0.99
	8.	age_group=(20, 30] ∩ on_thyroxine=F ∩ lithium=F ∩ goitre=F ==> class=health	0.99
	9.	age_group=(20, 30] ∩ on_thyroxine=F ∩ pregnant=F ==> class=health	0.99
	10.	age_group=(20, 30] ∩ on_thyroxine=F ∩ pregnant=F ∩ goitre=F ==> class=health	0.99
Sick	1.	age_group=(60, 70] ∩ sick=F ==> class=sick	1
	2.	age_group=(60, 70] ∩ on_antithyroid_medication=F ∩ sick=F ==> class=sick	1
	3.	age_group=(60, 70] ∩ sick=F ∩ goitre=F ==> class=sick	1
	4.	age_group=(60, 70] ∩ sick=F ∩ tumor=F ==> class=sick	1
	5.	age_group=(60, 70] ∩ on_thyroxine=F ∩ query_on_thyroxine=F ∩ thyroid_surgery=F ==> class=sick	1
	6.	age_group=(60, 70] ∩ on_thyroxine=F ∩ query_on_thyroxine=F ∩ query_hypothyroid=F ==> class=sick	1
	7.	age_group=(60, 70] ∩ on_thyroxine=F ∩ query_on_thyroxine=F ∩ lithium=F ==> class=sick	1
	8.	age_group=(60, 70] ∩ on_thyroxine=F ∩ pregnant=F ∩ thyroid_surgery=F ==> class=sick	1
	9.	age_group=(60, 70] ∩ on_thyroxine=F ∩ pregnant=F ∩ query_hypothyroid=F ==> class=sick	1
	10.	age_group=(60, 70] ∩ on_thyroxine=F ∩ pregnant=F ∩ lithium=F ==> class=sick	1

D. *ARB for disease diagnosis*

The objection of this subsection is to measure performance of ARB and make a comparison with other algorithms by the same thyroid disease dataset in different papers. The main performance of metric is the percentage of correct classification.

i. *Selection and expression of input*

Based on the above-mentioned association rules, we will filter the baseline attributes and delete two attributes which are not used from association rules. Then we integrate eight most frequent factors with five clinical test indicators (*TSH, T3, TT4, T4U and FTI*) as the attributes of ARB, and the baseline attributes with five clinical test indicators (*TSH, T3, TT4, T4U and FTI*) are as new baseline attributes in this subsection. The ARB attributes are shown in Table VII.

TABLE VII
ARB ATTRIBUTES OF THYROID-DATA DATASET

Attribute ID	Attribute Name	Attribute Explanation
1	*age*	Age interval
2	*sex*	M = Male or F = female
3	*on_thyroxine*	Whether taking thyroxine drugs T=true or F=False
4	*on_antithyroid_medication*	Whether taking anti-thyroid drugs T=true or F=False
5	*sick*	Whether sick T=true or F=False
6	*pregnant*	Whether pregnant T=true or F=False
7	*thyroid_surgery*	Whether in thyroid surgery T=true or F=False
8	*query_hypothyroid*	Whether had hypothyroidism T=true or F=False
9	*query_hyperthyroid*	Whether had hyperthyroidism T=true or F=False
10	*lithium*	Whether taking drugs containing lithium T=true or F=False
11	*tumor*	Whether have thyroid tumor T=true or F=False
12	*TSH*	Diagnosis indicators of thyroid disease
13	*T3*	Diagnosis indicators of thyroid disease
14	*TT4*	Diagnosis indicators of thyroid disease
15	*T4U*	Diagnosis indicators of thyroid disease
16	*FTI*	Diagnosis indicators of thyroid disease
17	*class*	Health or sick

ii. *Selection of base algorithm based on thyroid-data dataset*

The purpose of using bagging algorithm is to increase the accuracy of detecting and classifying thyroid disease, and base algorithm plays an essential role on its performance. Different base algorithms, which are trained in the same dataset, will produce different results. Therefore, in order to improve the classification accuracy of thyroid disease diagnosis, several common base algorithms (Naive Bayes, SMO, C4.5, C4.5 graft and KNN) are used to compare via K-fold cross-validation. K-fold cross-validation divides the dataset into approximately equal K subsets, one of which is used as test data in turn, and the remaining K-1 subset is used as training set. The average of K-fold results is used as the classification accuracy of the algorithm, and then the performance of each algorithm in thyroid disease datasets is observed.

In this subsection, baseline is the accuracy before attribute selection and the accuracy after attribute selection by association rule mining algorithm is as ARB. The analysis results are shown in Table VIII and Table IX. In Table VIII, 3-fold cross-validation is used to estimate the performance of the base algorithms. And it can be seen that C4.5 graft algorithm has the best accuracy (i.e. 96.9589), though association rules mining algorithm does not have an effect on it. Considering that all the other four algorithms have been affected by the attribute selection of Apriori algorithm, we can regard this situation as its high accuracy. The accuracy of SMO gets the minimum before selecting attitudes, which is 90.0152%. Furthermore, ARB has an adverse impact on it, which there is a 0.03% decrease nearly. We can ignore its decrease, since the amount of correctly classified instance only reduce by one, and its accuracy in Table IX has a significantly improvement. Table VIII shows that Naive Bayes has the most significant improvement, and it is nearly 0.2 percentages. KNN's performance is in the middle among five algorithms.

In Table IX, we apply 10-fold cross-validation to estimate base algorithms a second time. We can see that C4.5 graft also performs best and the accuracy of C4.5 algorithm is second only to C4.5 graft before attribute selection, and the accuracies are 97.4911 and 97.3898 respectively. After attribute selection, the accuracy of two algorithms both decreases, though there is a difference of only one incorrectly classified instance. And we deal with the same situation as above. In this table, KNN improves more significant than Naive Bayes, which the increase is more than 0.2%. And it performs also in the middle status.

In the experimental results of two tables, the accuracy of Naive Bayes and SMO algorithms are maintained at about 90%. SMO always achieves the least accuracy. It is probably because the attributes in thyroid dataset are associated, and SMO algorithm may not find maximum marginal hyperplane. Maybe the same reason cause Naive Bayes algorithm doesn't perform well, since the premise of Naive Bayes algorithm is that the attributes are independent of each other. The performance of KNN algorithm is always in the middle.

To summarize, we can see that C4.5 graft algorithm performs always better than C4.5. The C4.5 graft algorithm is an improvement of the C4.5 algorithm. In the literature [33], the C4.5 graft algorithm also performs well in many other datasets. Therefore, we chooses C4.5 graft algorithm as the base algorithm.

TABLE VIII

THE ACCURACY OF DIFFERENT BASE ALGORITHMS VIA 3-FOLD CROSS-VALIDATION

Algorithm (%)	Naive Bayes	SMO	C4.5	C4.5graft	KNN
Baseline	90.7248	90.0152	96.7562	96.9589	92.8535
ARB	90.9022	89.9899	96.7816	96.9589	92.9549

TABLE IX

THE ACCURACY OF DIFFERENT BASE ALGORITHMS VIA 10-FOLD CROSS-VALIDATION

Algorithm (%)	Naive Bayes	SMO	C4.5	C4.5graft	KNN
Baseline	90.4967	90.2680	97.3898	97.4911	93.3604
ARB	90.5981	90.3179	97.3644	97.4658	93.5884

iii. *Selection of ensemble algorithm based on thyroid-data dataset*

In this subsection, we compare bagging algorithm with other common ensemble algorithms (i.e. Adaboost, Random Forest and Rotation Forest) and only apply 10-fold cross-validation to analyze the classification performance. The experimental results are shown in Fig.5, Fig.6, and Table X. Results show that Adaboost algorithm not only need more time to establish the model, but also cannot achieve a satisfactory performance. Random Forest algorithm and Rotation Forest algorithm have the similar accuracy, and they are 97.2884% and 97.2631% respectively. But Random Forest algorithm performs better than Rotation Forest algorithm significantly in both aspects of modeling time and accuracy. Comprehensive analysis is conducted that bagging algorithm has best performance among all the ensemble algorithms.

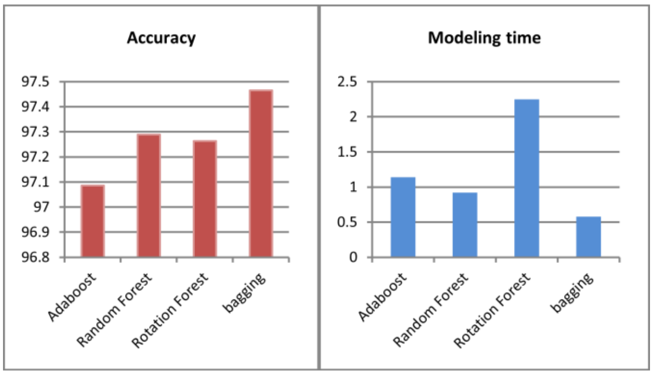

Fig.5.The accuracy of different ensemble algorithms

Fig.6.The mode different modeling time ensemble algorithms

TABLE X

THE COMPANION BETWEEN DIFFERENT ENSEMBLE ALGORITHMS

Algorithm	Adaboost	Random Forest	Rotation Forest	Bagging
Accuracy (%)	97.0857	97.2884	97.2631	97.4658
Modeling Time(s)	1.14	0.92	2.25	0.58

iv. *Experiment analysis based on new-thyroid dataset*

In this subsection, we apply new-thyroid dataset which is also in UCI machine learning repository to bagging algorithm. Then we compare ARB with algorithms in literature [30-32] via 3-fold cross-validation and 10-fold cross-validation.

The new-thyroid dataset contains three classes and 215 samples. These classes are assigned to the values that correspond to the hyper, hypo and normal function. All samples have five features. They are as follows.

1. T3-resin uptake test (A percentage).
2. Total serum thyroxin as measured by the isotopic displacement method.
3. Total serum triiodothyronine as measured by radioimmuno assay.
4. Basal thyroid-stimulating hormone (TSH) as measured by radioimmuno assay.
5. Maximal absolute difference of TSH value after injection of 200 mg of Thyrotropin-releasing hormone as compared to the basal value.

From Table XI, we can see that ARB has the best performance in 10-fold cross-validation (i.e. 94.8837%), but its accuracy is only 0.07% higher than PNN in paper [31]. Table XI also shows that ARB, which achieves an accuracy of 91.6279%, doesn’t perform best in 3-fold cross-validation, though it is only after PNN which is 94.43% and MLNN with LM which is 92.96%.

TABLE XI

THE COMPANION USED NEW-THYROID DATASET

Study	Method	Accuracy (%)
Paper[30]	MLP with bp(3*FC)	86.33
	MLP with fbp(3*FC)	89.80
	CSFNN(3*FC)	91.138
	RBF(3*FC)	79.08
Paper [31]	MLNN with LM(3*FC)	92.96
	PNN(3*FC)	94.43
	LVQ(3*FC)	89.79
	MLNN with LM(10*FC)	93.19
	PNN(10*FC)	94.81
	LVQ(10*FC)	90.05
Paper [32]	MLP neuronal function	94.0
This paper	ARB(3*FC)	91.6279
	ARB(10*FC)	94.8837

V. CONCLUSIONS AND FUTURE WORK

Attribute selection is a crucial step in disease classification diagnosis. We imitate the diagnosis process of traditional Chinese medicine, and use the association rule mining algorithm (i.e. Apriori algorithm). We can not only filter the useless attributes in order to reduce dimension, but also analyze the relationship among thyroid disease attributes. Then, according to the factor of gender, the rules are further mined and studied in more details. Through extracted rules, it is found that the risk of thyroid disease increases with age, and the elderly from 60 to 80 are most likely to suffer from thyroid disease. And if the elderly are sick, adequate prevention of thyroid disease should be done. Most of the thirties have more chances of being free from thyroid disease. And people who are at the two age intervals (30, 40] and (50, 60] will have more chances of the recurrence of thyroid disease, especially for men. In terms of gender, women have a greater chance than men. In the twenties, women have less risk. The aforementioned conclusions show that gender and age are two most important factors leading to thyroid disease. These are also supported by existing clinical medical research. In future experimental research, gender and age should be listed as important factors impacting thyroid disease.

Then, we compare different classification algorithms in thyroid-data dataset to choose a base algorithm which has the best performance. And then we compare various ensemble algorithms with bagging algorithm. The results show that bagging algorithm has best performance than the others.

Finally, we use the framework ARB to diagnose thyroid disease based on new-thyroid dataset. By Comparing with other authors' previous work, ARB always has a better performance. And it also preliminarily illustrates the feasibility and practical value of the framework ARB in medical aided diagnosis.

ACKNOWLEDGMENTS

This work was supported by General Scientific Research Projects of Liaoning Province (2019LNJC07) and University of Science and technology Liaoning Talent Project Grants.

REFERENCES

[1] Wang Xuemei, Li Jiaru and Wang Liping, "Analysis of clinical cases of thyroid diseases," *Chinese Journal of Laboratory Diagnosis*, vol. 16, no. 1, pp.127-129, 2012

[2] Liu Guozhong, Liu Hui and Zhao Peng, "Physical Education Research on Human Age Segmentation," *Science and Education Collection*, vol. 14, no. 22, pp.150-150, 2013

[3] Tian Hui, "The prevalence and influencing factors of thyroid diseases in China," *Chinese Journal of Multiple Organ Diseases in the Elderly*, vol. 12, no. 2, pp.81-84, 2013

[4] P. N. Taylor, D. Albrecht, A. Scholz, G. Gutierrez-Buey, J. H. Lazarus, C. M. Dayan and O. E. Okosieme, "Global epidemiology of hyperthyroidism and hypothyroidism," *Nature Reviews Endocrinology*, vol. 14, no. 5, pp.301–316, 2018

[5] M. Zhan, G. Chen, C. M. Pan and et al, "Genome-wide association study identifies a novel susceptibility gene for serum TSH levels in Chinese populations," *Human Molecular Genetics*, vol. 23, no. 20, pp.5505-5517, 2014

[6] Kaloumenou, L. Duntas, M. Alevizaki, E. Mantzou, D. Chiotis, C. Mengreli, I. Papassotiriou, G. Mastorakos, C. Dacou-Voutetakis, I. Kaloumenou, L. Duntas, M. Alevizaki, E. Mantzou, D. Chiotis, C. Mengreli, …C. Dacou-Voutetakis, "Gender, Age, Puberty, and BMI Related Changes of TSH and Thyroid Hormones in Schoolchildren Living in a Long-standing Iodine Replete Area," *Hormone and Metabolic Research*, vol. 42, no. 04, pp.285–289, 2010

[7] M. Bauer, T. Glenn, M. Pilhatsch, A. Pfennig and P. C. Whybrow, "Gender differences in thyroid system function: relevance to bipolar disorder and its treatment," *Bipolar Disorders*, vol. 16, no. 1, pp.58–71, 2013

[8] Wang Mingxue, Yang Biwei and Zhang Hua, "Necessity of establishing reference intervals for five indicators of thyroid function by age and sex in laboratory," *International Journal of Laboratory Medicine*, vol. 40, no. 04, pp.464-468, 2019

[9] W. Gulbinat. (1997). What is the role of WHO as an intergovernmental organization in the coordination of telematics in health care? Available: https://www.hon.ch/Library/papers/gulbinat.html

[10] Samo Riyanarto, Dewandono Rahadian Dustrial, Ahmad Tohari, Naufal Mohammad Farid and Sinaga Fernandez, "Hybrid association rule learning and process mining for fraud detection," *IAENG International Journal of Computer Science*, vol. 42, no. 2, pp.59-72, 2015

[11] Ogunde, O. Adewale , Folorunso, Olusegun, Sodiya and S. Adesina, "The design of an adaptive incremental association rule mining system," in *Lecture Notes in Engineering and Computer Science, WCE 2015 - World Congress on Engineering 2015*, pp.172-177, 2015

[12] Kaoungku Nuntawut, Suksut, Keerachart, Chanklan Ratiporn Kerdprasop Kittisak and Kerdprasop Nittaya, "Data Classification Based on Feature Selection with Association Rule Mining," in *Lecture Notes in Engineering and Computer Science, Proceedings of the International MultiConference of Engineers and Computer Scientists 2017*, pp.321-326, 2017

[13] Wang Yingquan, Murata Tomohiro, "Association Rule Mining with Data Item including Independency based on Enhanced Confidence Factor," in *Lecture Notes in Engineering and Computer Science, Proceedings of the International MultiConference of Engineers and Computer Scientists 2017*, pp.359-363, 2017

[14] Z. Yang, L. Lei and Jiancheng, "Study on Distribution Rules of TCM Signs and Symptoms and Syndrome Elements in Essential Hypertension Based on Data Mining," *Chinese Journal of Information on Traditional Chinese Medicine*, 2019

[15] Gao Bing, Wang Jian, Guo Jinchen, Cheng Yue, Huang Hui, Zong Yanping, Feng Ye and Zhang Hao, "Construction of mathematical model of Xinan Wangs internal medicine in treatment of epigastric pain based on data mining technology," *Chinese Traditional and Herbal Drugs*, vol. 49, no. 23, pp.5705-5711, 2018

[16] Zhu XL and et al, "Research on Classification of Tibetan Medical Syndrome in Chronic Atrophic Gastritis," *APPLIED SCIENCES-BASEL*, vol. 9, no. 8, 2019

[17] Dong Wenzhe, Liu Jian; Xin Ling, Fang Yanyan and Wen Jianting, "Data mining research on different formulations of Tripterygium wilfordii ameliorating immune inflammation in patients with rheumatoid arthritis," *Immunological Journal*, vol. 34, no. 10, pp.894-899,905, 2018

[18] Yu Ling, Wang Yingxiao and Li Qizhong, "Preliminary study on medication characteristics of internal treatment of DING Gan-ren's medical records in external medicine," *China Journal of Traditional Chinese Medicine and Pharmacy*, vol. 38, no. 8, pp.3551-3553, 2018

[19] Lv Y, Peng S, Yuan Y and et al, "A classifier using online bagging ensemble method for big data stream learning," *Tsinghua Science and Technology*, vol. 24, no. 4, pp.379-388, 2019

[20] Liu Hailing and Zang Xian, "Decision tree and bagging algorithm for the automatic identification of epithelial cell of wound," in *1st International Conference on Advanced Algorithms and Control Engineering, ICAACE 2018*, pp.1087, 2018

[21] Dai Peng, Gwadry-Sridhar Femida and Bauer Michael, "Bagging ensembles for the diagnosis and prognostication of Alzheimer's disease," in *30th AAAI Conference on Artificial Intelligence, AAAI 2016*, pp.20-23, 2016

[22] Prata Marco, Peixoto Hugo, MacHado José and Abelha António, "Data mining in urgency department: Medical specialty discharge prediction," in *16th International Industrial Simulation Conference 2018*, pp.28-35, 2018

[23] R. Gart and Guo Zhiling, "Classification of Thyroid Diseases," *Oncology and Translational Medicine (English)*, no. 4, pp.191-192, 1993

[24] Han Jiawei and Micheline Kamber, Data Mining Concept and Algorithm, Fan Ming, Translated by Meng Xiaofeng. Beijing, Machinery Industry Press, 2001

[25] Jiaweihan, Micheline Kamber, JianPei and et al, Data Mining: Concept and Algorithm, Machinery Industry Press, 2012

[26] R. Agrawal, T. Imieliński and A. Swami, "Mining association rules between sets of items in large databases," in *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data - SIGMOD '93*, 1993

[27] Yao Xusheng, Yang Jing, Xie Yingfu and He Jianfeng, “Application of Association rule mining algorithms in Clinical Medical Diagnosis,” *Software Guide*, vol. 17, no. 3, pp.162-164, 2018
[28] Cai Hong, Chen Hui and Chen Bo, “Research on Optimization Algorithm of minimum support threshold setting for association rule mining,” *Microcomputer application*, vol. 27, no. 6, pp.33-36, 2011
[29] Zhang Wenxi, Su Haixia, Shang Lei, Sun Lijun and Zhang Yuhai, “Comparative study on predicting the progression of Alzheimer's disease based on BP neural network and RBF neural network,” *Progress in Modern Biomedicine*, vol. 17, no. 4, pp.738-741, 2017
[30] L. Ozyilmaz and T. Yildirim, “Diagnosis of thyroid disease using artificial neural network methods,” in *Proceedings of the 9th International Conference on Neural Information Processing 2002*, pp.2033-2036, 2002
[31] Temurtas, “A comparative study on thyroid disease diagnosis using neural networks,” *Expert Systems with Applications*, vol. 36, no. 1, pp.944-949, 2009
[32] S. Isa, Z. Saad, S. Omar and et al, “Suitable MLP network activation functions for breast cancer and thyroid disease detection,” in *Proceedings of the 2nd International Conference on Computational Intelligence, Modelling and Simulation 2010*, pp.39-44, 2010
[33] I. Webb, “Decision tree grafting from the all tests but one Partition,” in *Proceedings of the 16th International Joint Conference on Artificial Intelligence 1999*, pp.702-707, 1999