FF-Based Feature Selection for Improved Classification of Medical Data

YAN WANG, LIZHUANG MA
Department of Computer Science & Engineering
Shanghai Jiao Tong University
No. 1954, Huashan Road, Shanghai
P.R.CHINA
wangyan8383@sjtu.edu.cn http://www.sjtu.edu.cn

Abstract: - In processing the medical data, choosing the optimal subset of features is important, not only to reduce the processing cost but also to improve the classification performance of the model built from the selected data. Rough Set method has been recognized to be one of the powerful tools in the medical feature selection. However, the high storage space and the time-consuming computation restrict its application. In this paper, we propose two new concepts: *discernibility string* and *feature forest*, and an efficient algorithm, the Feature Forest Based (FF-Based) algorithm, for generation of all reducts of a medical dataset. The algorithm consists of two phases: feature forest construction phase and disjunctive normal form computation phase. In the first phase, the discernibility strings that are the concatenation of some of features between two different cases construct the feature forest. In the second phase, the disjunctive normal form is computed to reduct features based on feature forest. The experimental results on the medical datasets of UCI machine learning repository and a real liver cirrhosis dataset show that the algorithms of this paper can efficiently reduce storage cost and improve the classification performance.

Key-Words: - Feature selection, rough set, disjunctive normal form, feature forest, discernibility string

1 Introduction

Many factors affect the success of machine learning on the medical datasets. The quality of the data is one such factor. If information is irrelevant and redundant or the data is noisy and unreliable then knowledge discovery during training is more difficult. Feature selection is the process of identifying and removing as much of the irrelevant and redundant features as possible.

In processing medical data, the advantage of choosing the critical features is as follows:

- Simplifying data description may facilitate physicians to make a sound and prompt diagnosis;
- Having fewer features means that less data need to be collected, as we know; collecting data is never an easy job in medical applications because it is time-consuming and costly.

In recent decades, considerable research efforts have been devoted to using data mining techniques to discover useful medical knowledge and rules automatically.

Hongmei Yan et al. proposed a real-coded genetic method to select critical features essential to the heart diseases diagnosis. In result, 24 critical features have been identified, and their corresponding diagnosis weights for each heart disease of interest have been determined. The critical diagnostic features and their clinic meanings are in sound agreement with those used by the physicians in making their clinic decisions [1]. Tsang-Hsiang Cheng et al. adopted correlation feature selection (CFS) method to obtain the feature subset about cardiovascular disease. After selecting the feature subsets, the predictive power of a classifier on the majority class was improved [2]. R.E. Abdel-Aal used the group method of data handling (GMDH) to reduct 22% and 54% dimensionality for the breast cancer and heart disease data, respectively, lead to the improvements in the overall classification performance [3].

Among these techniques, rough set method has been recognized to be particularly powerful in medical knowledge discovery. Rough set theory (RST) was proposed by Pawlak [4], which was a valid mathematical theory to deal with imprecise, uncertain information and was to obtained as simple as rules from the given data base by reducing the data base while holding the original degree of consistency [8,13]. Liu Yaohe et al. extracted the concise and intelligible classification rules of medical sample data based on rough set theory [5]. Qin Zhongguang developed a traditional Chinese

medicine rheumatic arthritic intelligent diagnosis system to extract useful diagnosis knowledge based on rough set method. The results showed that the diagnostic accuracy of Rough set for rheumatoid arthritis was greatly higher than that of fuzzy set [6]. Tsumoto proposed a rough set algorithm to generate diagnostic rules based on the hierarchical structure of differential medical diagnosis. The induced rules can correctly represent experts' decision processes [7]. Based on the rough set method, Xiangyang Wang et al. proposed attribute reduction algorithm that employed a search method based on particle swarm optimization (PSO); the reducts found by the proposed algorithm were more efficient and could generate decision rules with better classification performance [18]. Bazan compared the rough set-based methods, in particular dynamic reducts, with statistical method, neural network, decision tree and decision rule. He analyzed the medical data, i.e. lymphography, breast cancer and primary tumors, and found that error rates for rough sets are fully comparable as well as often significantly lower than that for other techniques [9].

Rough set theory has two fundamental concepts to deal with these problems: feature selection and core. There are many methods available for core and feature selection based on rough set [5-8, 10, 16-17]. Among them, methods based on discernibility matrix are of considerable benefits [5-6, 10], because each entry $m_{i,j}$ of the matrix corresponding to objects x_i and x_j includes the conditional features in which the two objects' values differ. The methods based on discernibility matrix are concise and efficient, but the matrix occupies high storage space and the computation of final disjunctive normal form (DNF) is time-consuming.

At present, most researches are focused on reducing the storage space of the existing feature selection methods and improving the efficient of DNF computation based on discernibility matrix [6, 10-11]. Some efforts have been made [12, 14]. However, the time complexity is still $O(n^2(|C|^{n^2}+|D|))$ [14], where |C| is the number of conditional feature and |D| is the number of decision feature.

In this paper, we proposed a two-phase feature selection approach called as *Feature Forest* Based method to discover significant feature sets from a given database table. The method can efficiently reduce the cost of storage space and improve the DNF computation. This method includes two phases-feature forest construction phase and disjunctive normal form (DNF) computation phase. Algorithm of this paper is experimental using some standard medical datasets and a real liver cirrhosis dataset for testing both time and space complexities. Experimental results show that the algorithms of this paper can efficiently reduce storage cost and be computationally inexpensive based on these datasets.

The other parts of this paper are organized as follows. Section 2 reviews some related theories and defines related symbols about rough set method. Section 3 introduces some corresponding definitions, proposes the framework of the FF-Based algorithm, and gives an example about the algorithm. Section 4 gives the experimental results through extensive experiments. Section 5 summarizes this paper and discusses future works.

2 Basic Rough Set Theory

Rough set theory has been introduced by Pawlak [4] to deal with imprecise or vague concepts. In recent years, we witnessed a rapid growth of interest in rough set theory and its applications, worldwide. Here, we introduce only the basic notation from rough set approach used in the paper.

A decision table is denoted as $S=(U,A,V,f)$, where $U=\{x_1,x_2,...,x_n\}$ is called universe, which is a non-empty finite set of cases. A denotes the set of m features in U, $A=C\cup D$, where C is the set of condition features and D is the set of decision features.

In the medical data, $C=\{C_1,C_2,...C_i,...,C_{m-1}\}$, where C_i is often the ordinal value such as 1,2 or 0,1 that represent the case have some symptoms or not; $D=\{d\}$ is often a singleton set, where d is the decision feature that denotes the class labels of cases. $V=\bigcup_{C_i\in A}V_{C_i}$, where V_{C_i} is the domain of feature C_i.

$f:U\times A\to V$ is the information function such that $\forall C_i\in A, x_i\in U, f(x_i,C_i)\in V_{C_i}$. Other formal definitions about rough set can be found in [15].

In the consistent decision system, that is, if the conditional features of two or more cases are identical, the decision features are the same; the discernibility matrix [5, 12] is defined as a $n\times n$ matrix of S with every element given by

$$\alpha(x_i,x_j)=\begin{cases}C_k\in C\mid f(x_i,C_k)\neq f(x_j,C_k) & f(x_i,D)\neq f(x_j,D)\\ 0 & f(x_i,D)=f(x_j,D)\end{cases} \quad (1)$$

where $1\le k\le m-1, 1\le i\le n, 1\le j\le n$.

The corresponding discernibility function for the decision table S is a Boolean function defined by:

$$\Delta = \prod_{(x_i,x_j)\in U\times U} \sum \alpha(x_i,x_j) \quad (2)$$

where $\sum \alpha(x_i,x_j)$ denotes $C_1 \vee C_2 \vee ... \vee C_k$.

The discernibility function describes constraints, which should be preserved if one would like to preserve discernibility between all pairs of discernible objects from S.

We form the discernibility function by the conjunction of all $\sum \alpha(x_i,x_j)$ together. The final DNF obtained by simplifying the function Δ yields possible reducts for the given dataset.

Unfortunately, the final feature reduction based on the discernibility function needs compute $n^2(|C|^{n^2}+|D|)$ [14] times conjunctive normal form approximately. In order to reduce the computation times of basic discernibility matrix, in this paper, we introduce a *Feature Forest* Based method in Section 3 to compute all the reducts for the medical dataset.

3 Proposed Feature Reduct Algorithm

Based on the characteristic of the medical dataset, this paper assumes:

- Every decision system has one decision feature.
- Every decision system is the consistent decision system, that is, if the class labels of two or more cases are identical, they may be the same disease.

Based on the above assumptions, we proposed a feature reduction approach called as *Feature Forest* Based method to reduct features of the medical dataset. The proposed approach is shown in Figure 1. There are two phases in the proposed method-feature forest construction phase and DNF computation phase.

In the feature forest construction phase, the given dataset is transformed into a decision forest with the decision feature as root of each tree. We compute the discernibility string between the leaf of the left tree and that of the right tree. In the DNF computation phase, a set of relevant and enough features are selected and used to present the dataset. The details of the two phases are described in following sub-sections.

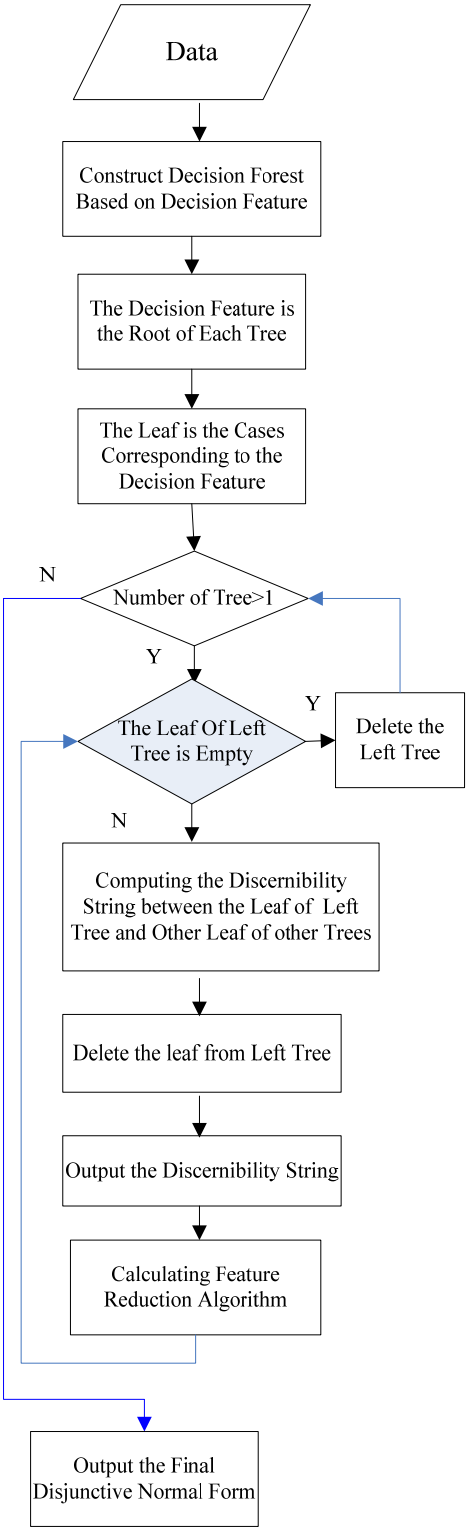

Fig. 1 The proposed method

3.1 Problem definitions

In this section, we define the symbols which have been used in our algorithm. The domain set of decision feature is represented as $\{d_1, d_2, ..., d_k\}$, where each element is a possible value of decision feature and k is the number of possible values of decision feature. Let R_j denote the set of records with the corresponding decision feature d_j, R_j can then be represented as $\{R_{j(1)}, R_{j(2)}, ..., R_{j(t)}\}$, where $R_{j(i)}$ is the i^{th} case in R_j. Let $V_C(ji)$ denote the domain of C in record $R_{j(i)}$. $V_C(ji)$ can then be represented as $\{ V_{C_1}(ji), V_{C_2}(ji), ..., V_{C_{m-1}}(ji) \}$, where each element is a possible value of C_j.

Table 1 shows an example about pneumonia and phthisis borrowed from [5]. The storage space of the basic discernibility matrix is $20 \times 20 \times 4$. The target table U has twenty records and five features $C = \{C_1, C_2, C_3, C_4\}$, C_1, C_2, C_3, C_4 are the conditional features whose meanings shown in Table 2, $D = \{d_1, d_2\} = \{1, 2\}$ is the domain of the decision feature.

Table 1 An example of a target table

U	C1	C2	C3	C4	D	U	C1	C2	C3	C4	D	U	C1	C2	C3	C4	D	U	C1	C2	C3	C4	D
1	4	3	1	3	1	6	2	1	3	1	2	11	3	1	2	3	2	16	4	3	2	3	1
2	3	3	1	3	1	7	4	2	1	3	1	12	2	2	1	3	2	17	3	1	2	2	2
3	1	1	3	1	2	8	3	1	1	3	1	13	1	2	3	1	2	18	1	3	2	1	2
4	3	1	2	1	1	9	3	2	1	3	1	14	3	2	2	1	1	19	2	1	4	1	2
5	4	3	4	2	2	10	4	3	2	1	1	15	4	2	2	3	1	20	4	3	3	2	2

Table 2 Description of features in example

No.	Label	Features	Description
1	C1	1,2,3,4	1: no fever; 2: low fever; 3: middle fever;4: high fever
2	C2	1,2,3	1: mild cough; 2: middle cough; 3: acute cough
3	C3	1,2,3,4	1: sheet; 2: punctuation; 3: funicular; 4: amphoric
4	C4	1,2,3	1: normal; 2: dry rales; 3: bubbles
5	D	1,2	1: pneumonia; 2: phthisis

3.2 Feature forest construction phase

Definition 1 (*leaf vector*). In the medical dataset, $V_{C_k}(ji)$ is often an ordinal value. A *leaf vector* is defined as the combination of the bit string $V_{C1}(ji)V_{C2}(ji)...V_{C(m-1)}(ji)$, which is used to keep the information of a case.

Definition 2 (d*ecision forest*). A decision forest DT constructed by each R_j. In each R_j tree, d_j is the root and the leaves are formed by all of the cases with the same decision features d_j. The tree with less leaves is arranged on the left of DT.

Example 1. Based on the above definition, in this phase, the target table is first transformed into a *decision forest* DT. For instance for Table 1, the domain of D is {1,2}, DT has two trees. R_1 has ten cases with the decision features *1* while other ten cases are in R_2. We arrange R_1 in the left tree, the value of the root is *1* and the first *leaf vector* is 4313. The *decision forest* of Table 1 is shown in figure 2 and figure 3 respectively.

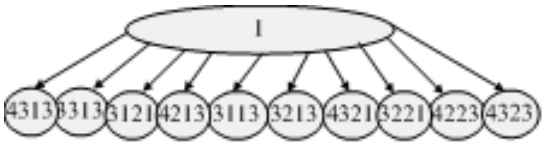

Fig. 2 The decision forest R_1 from Table 1

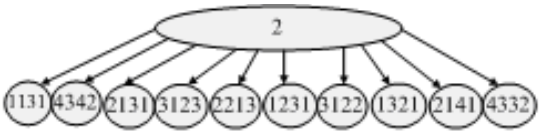

Fig. 3 The decision forest R_2 from Table 1

Definition 3 (*discernibility string*). A discernibility string DS is the concatenation of C_j such as $C_pC_q...C_l$ keeping the order in accord with your need. In this paper, the order is the original conditional feature order of C.

The discernibility string is constructed by two ways. The first way is using $V_C(ji) \oplus V_C(kt)$ to obtain the DS, where $k \neq j$. Here, the $\oplus$ bitwise operator is used for distinguishing each C_j between two cases. The second way is formed by $DS_1 \wedge DS_2$, the $\wedge$ is a **AND** operator.

Example 2. The first leaf vector in R_1 is 4313, a leaf in R_2 is 1131, computing $4313 \oplus 1131$, because $4 \neq 1, 3 \neq 1, 1 \neq 3, 3 \neq 1$, the corresponding *discernibility string* DS is $C_1C_2C_3C_4$ which is the input of the DNF computation algorithm.

Example 3. If $DS_1 = C_1, DS_2 = C_3$, then $DS_1 \wedge DS_2 \rightarrow C_1C_3$

The discernibility string formed by the second way is to form the leaf of feature forest, which will be defined in 3.3.

3.3 DNF computation phase

In this phase, the computable course of the DNF is described. The related theorem and definition is as follows.

Definition 4 (*feature forest***)**. A feature forest FF is the combination of feature tree FT_j. The root of each tree denotes the frequency of C_j. $CF_j(k)$ which is a DS denotes the k[th] leaf in FT_j.

Theorem 1. Conjunctive normal form and disjunctive normal form follow associative law and commutative law. That is,

$((C_1 \wedge (C_2 \vee C_3)) \wedge (C_3 \vee C_4)) =$
$((C_1 \wedge (C_3 \vee C_4)) \wedge (C_2 \vee C_3))$

Theorem 2. If $DS_1 \subseteq DS_2$ then $DS_1 \wedge DS_2 = DS_1$.

From theorem 1 and 2, we can deduce that when computing $DS_i \wedge (DS_1 \vee ...DS_j ... \vee DS_k)$

If $DS_i \subset DS_j$ then

$DS_i \wedge (DS_1 \vee ...DS_j ... \vee DS_k) = DS_i$;

If $DS_i \supset DS_j$ and

$DS_i \cap DS_p = \varnothing, where \ \ 1 \leq p \leq k \ and \ p \neq j$

then $DS_i \wedge (DS_1 \vee ...DS_j ... \vee DS_k) = DS_j$

Based on the definitions and the theorems, some discernibility strings need not be concerned with the DNF computation. The proposed algorithm checks if the discernibility string is discarded or not at first. Therefore, the efficiency of DNF computation is improved. The detail about the algorithm is as follows:

Algorithm 1. DNF computation

Input: discernibility string DS

Output : each leaf $CF_j(k)$ of feature forest

Step 1: Constructing feature forest. The root of each FT_j is initialized with zero. If C_j appears in DS, the root of FT_j increases 1.

Step 2: C_j included in DS becomes the leaf node of FT_j.

Step 3: For (j=1 to $|C|$-1,k=1 to FT_j)

Step 3.1: If DS and a leaf of FT_j intersect at $CF_j(k)$.

Step 3.1.1: If $CF_j(k)$ is a singleton feature, where $1 < v$. C_j is deleted from DS.

Step 3.2: If DS is null, the leaf of FT_i is deleted, where $i \neq j$.

Step 3.3: If DS isn't null, concatenating each C_j with every leaf of FT_i which leaf is not empty, where $i \neq j, 1 \leq i \leq m-1$.

Step 3.4: If DS and any laves of FT_j without intersection. Concatenating C_j that is in DS with every leaf of FT_i which leaf is not empty to construct a new leaf , where $i \neq j, 1 \leq i \leq m-1$.

Step 4: Output all of the leaves of FT_j to construct the final feature set reduction.

3.4 An example

In this section, a simple example is given to demonstrate the proposed FF-Based algorithm, which generates three critical features from 20 cases with four conditional features. The full dataset is

shown in Table 1. At first $R_{1(1)}$ orders all of the leaf of R_2 and obtains the corresponding DS, and then the detail step is as follows:

1. The feature set is C_1, C_2, C_3 and C_4. We construct the feature forest with four trees. Each tree is in accord with a feature. At first, the value of root is zero. Figure 4 shows the result.

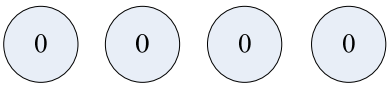

Fig. 4 Feature forest from Table 1

2. R_{11} is 4313 and R_{21} is 1131. Computing 4313 $\oplus$ 1131 to obtain the discernibility string $C_1C_2C_3C_4$. Add each C_j into FT_j, *where* $1 \le j \le 4$. Figure 5 shows the feature forest includes the leaf.

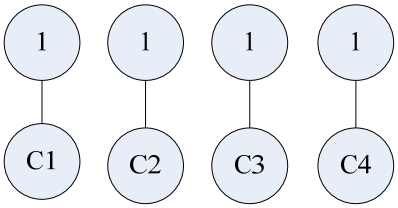

Fig. 5 Feature forest for $C_1C_2C_3C_4$

3. R_{11} is 4313 and R_{22} is 4342. Computing 4313 $\oplus$ 4342 to obtain the discernibility string C_3C_4 .The leaf of FT_3, FT_4 is the singleton feature, deleting C_3 and C_4 from DS. DS is null, we delete C_1, C_2 from FT_1, FT_2. Figure 6 shows the corresponding feature forest.

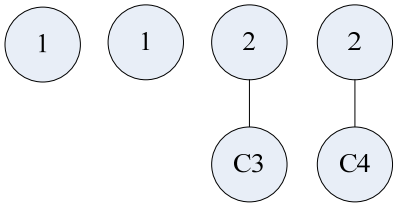

Fig.6 Feature forest for C_3C_4

4. R_{11} is 4313 and R_{23} is 2131. 4313 $\oplus$ 2131 $\rightarrow C_1C_2C_3C_4$. The leaf of FT_3, FT_4 is singleton feature, deleting C_3 and C_4 from DS. DS is changed into C_1C_2. Because FT_1 and FT_2 is empty, no new leaf is constructed. The feature forest is converted as Figure 7.

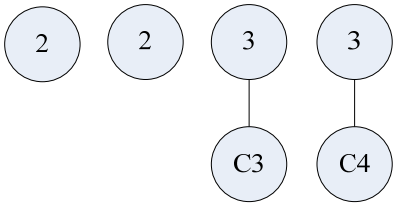

Fig. 7 Feature forest for corresponding $C_1C_2C_3C_4$

5. R_{11} is 4313 and R_{24} is 3123. 4313 $\oplus$ 3123 $\rightarrow C_1C_2C_3$. The leaf of FT_3 is the singleton feature, deleting C_3 from DS. DS is changed into C_1C_2. Concatenating C_1, C_2 with every leaf of FT_4 to obtain two leaf vectors C_4C_1, C_4C_2. Sorting the leaf vector according to the feature order and the leaf vector is changed into C_1C_4, C_2C_4. Adding the leaf vector to FT_1 and FT_2. The forest is shown in Figure 8.

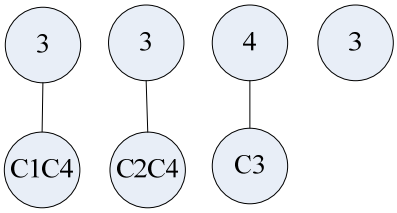

Fig.8 Feature forest for $C_1C_2C_3$

6. R_{11} is 4313 and R_{25} is 2213.4313 $\oplus$ 2213 $\rightarrow C_1C_2$. The leaf of FT_1 and FT_2 is not a singleton feature. Concatenating C_1, C_2 with every leaf vector of FT_3 to obtain C_1C_3, C_2C_3 and add them to FT_1 and FT_2. The forest is shown in Figure 9.

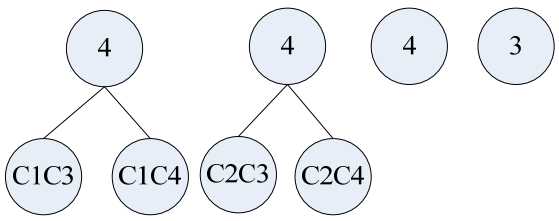

Fig. 9 Feature forest for C_1C_2

7. R_{11} is 4313 and R_{26} is 1231.4313 $\oplus 1231 \rightarrow C_1C_2C_3C_4$. The leaf of FT_1 and FT_2 is not a singleton feature and the leaf node of FT_3, FT_4 is empty. No new leaf vector is constructed. The forest is shown in Figure 10.

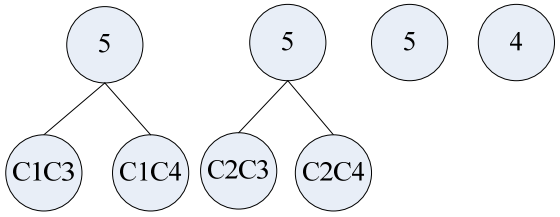

Fig. 10 Feature forest for $C_1C_2C_3C_4$

8. R_{11} is 4313 and R_{27} is 3122. 4313 $\oplus 3122 \rightarrow C_1C_2C_3C_4$. The situation is similar to step 7. Figure 11 shows the result.

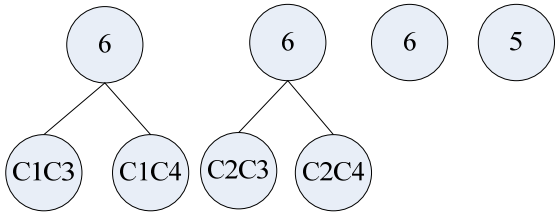

Fig.11 Feature forest for $C_1C_2C_3C_4$

9. R_{11} is 4313 and R_{28} is 1321. $4313 \oplus 1321 \rightarrow C_1C_3C_4$. The leaf of FT_1 is not a singleton feature and FT_2 has two leaf vectors. But two leaf vectors of FT_2 have intersection with DS. No new leaf vector is constructed. Fig. 12 shows the result.

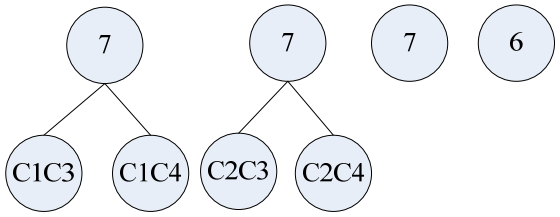

Fig.12 Feature forest for $C_1C_2C_3C_4$

10. R_{11} is 4313 and R_{29} is 2141. 4313 $\oplus 2141 \rightarrow C_1C_2C_3C_4$. The situation is similar to step 9.

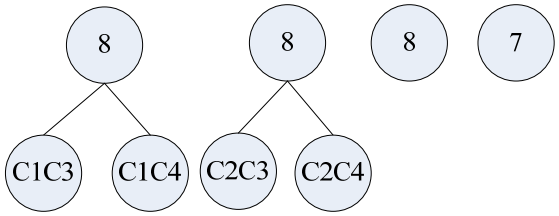

Fig.13 Feature forest for $C_1C_2C_3C_4$

11. R_{11} is 4313 and $R_{2(10)}$ is 4332. $4313 \oplus 4332 \rightarrow$is C_3C_4. Two leaf vectors of FT_1 and FT_2 have intersection with CF_{10}.

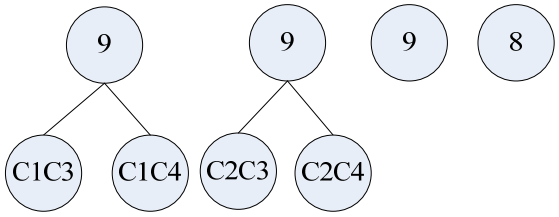

Fig.14 Feature forest for C_3C_4

12. The final forest for Table 1 is as follows:

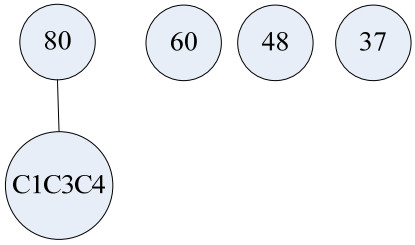

Fig.15 Feature reduction result for Table 1

After the above method is executed, we get the feature reduction and the frequency of features is obtained from the value of root at the same time. Combination the feature frequency and the length of feature string can find the desired feature reduction set. For Table 1, the final reduction is $C_1C_3C_4$.

4 Experimental Results

4.1 Description of datasets

In this paper, we use four datasets to verify the performance of FF-Based method. The first one is a real liver cirrhosis dataset. Based on the criteria for case-included and case-excluded, 268 cases with three different liver cirrhosis syndromes (i.e. stasis-heat smoldering syndrome, damp-heat smoldering

syndrome and liver-kidney yin deficiency syndrome) have been offered by Shanghai University of Traditional Chinese Medicine to constitute the sample dataset.

The dataset includes 85 cases with stasis-heat smoldering syndrome, 103 cases with damp-heat smoldering syndrome and 80 cases with liver-kidney yin deficiency syndrome.

Each case includes 68 recorded features, which are regarded as the basic symptoms required by physicians to identify the liver cirrhosis syndrome in clinic. Among all features, 40 are symptoms of Traditional Chinese Medicine such as lassitude and fatigue, night sweat, etc., the other 27 are signs such as pale tongue etc. and the last one is the syndrome label.

The second SPECT heart dataset is a public medical dataset that is downloaded at http://www.ics.uci.edu. The dataset describes the diagnosing of cardiac Single Proton Emission Computed Tomography (SPECT) images. Each of the patients is classified into two categories: normal and abnormal. The dataset has 267 cases that are described by 22 condition features and 1 decision feature.

The third Lung Cancer dataset is also downloaded at http://www.ics.uci.edu. It describes three types of pathological lung cancers. The dataset has 32 cases and each case has 56 conditional features, 1 decision feature.

The fourth dataset is shown in Table 1. It describes twenty pneumonia and phthisis cases.

4.2 Experimental results

Based on FF-Based algorithm, the experimental results are shown in Table 3, where $N1(DM)$ the number of entries for the discernibility matrix, $N2(DM)$ the number of entries for disjunctive normal form based on discernibility matrix without removing supersets without removing supersets. $N1(FF)$ the number of entries for the decision matrix, $N2(FF)$ the number of entries for disjunctive normal form based on feature forest method before removing supersets. $T2(FF)$ is the running time of feature forest algorithm, its unit is second and keeps three decimal digits. $String(FF)$ is the reduction result with the shortest length.

From Table 3, consistent with our expectation, the proposed approach occupies smaller storage space and less time-consuming than the original rough set method.

$String(FF)$ shows the critical features about the corresponding disease. The result is agreement with those used by the physicians in making their clinic decisions.

Table 4 compares the classification accuracy of the whole feature set with the feature subset from Table 3.After reducing feature, the classification accuracies of three typical classifiers are better than that of the original method. It indicates that the classification performance can be improved through the FF-Based feature selection algorithm.

Table 3 Experimental results

Datasets	DNF based on DM		DNF based on FF			
	N1(DM)	N2(DM)	N1(FF)	N2(FF)	T2(FF)	String(FF)
Pneumonia and Phthisis	100	2^{100}	20	C_4^2	0	C1, C3, C4
Lung Cancer	207	28^{207}	32	C_{56}^{28}	59.245	C39, C43, C52, C53, C55
Heart	11660	11^{11660}	267	C_{22}^{11}	0.078	C1, C3, C4, C5, C6, C7, C8, C9, C10, C13, C14, C16, C19, C20, C21, C22
Liver cirrhosis	12875	33^{12875}	228	C_{67}^{33}	103.412	C35, C38, C39, C45, C49, C51, C54, C55, C56, C61, C63, C64, C65, C66, C67

Table 4 Comparison of classification accuracy (%)

Datasets	Whole feature set			Reduction feature set from Table 3		
	BayesNet	RBFNetwork	SVM	BayesNet	RBFNetwork	SVM
Pneumonia and Phthisis	85	70	90	85	75	95
Lung Cancer	78.125	71.875	65.625	78.125	78.125	68.75
Heart	78.6517	82.0225	80.5234	80.5234	82.0225	81.2734
Liver cirrhosis	79.8507	75	76.4925	82.4627	79.8507	82.4627

5 Conclusions and Future Work

FF-Based method is proposed and its algorithm is carried out in this paper. The traditional rough set approach has two disadvantages that are time-consuming and large storage space. The new method decreases the storage space and improves the classification performance than the classical rough set method. The method contains the following characteristics:

(1) The dataset is transformed into a decision forest according to the value of the decision feature and the tree with less leaves is arranged at the left. Therefore, the storage space is the number of the leaves of the left tree.
(2) The discernibility string, which is obtained with the comparison of two cases, forms the feature forest. During the construction of the feature forest, some special discernibility strings according to **Theorem 2** are ignored. Therefore, the computation time of DNF shortens.
(3) Through experiments, the classification performance of typical classifiers is improved. We can conclude that FF-Based method fits feature selection on medical dataset.
(4) The feature selection problem is generally an NP-complete problem. Although, the proposed approach can process a larger amount of features than the traditional rough set approach, it still becomes unmanageable especially when the number of features is huge. In the future, we will continuously investigate and design efficient approaches to manage huge amounts of features.

Acknowledgement

We gratefully acknowledge all the researchers from the Shanghai Traditional Chinese Medicine University for the TCM databases and discussion of TCM topics. This research is partly supported by the traditional Chinese medicine etiologic study on the theory of insufficiency and damage causing stasis and blockage in liver cirrhosis of China 973 project (No. 2006CB504801) and by a grant from the Ph.D. Programs Foundation of Ministry of Education of China (20050248046).

References:

[1] Hongmei Yan, Jun Zheng, Yingtao Jiang, et al., Selecting critical clinical features for heart diseases diagnosis with a real-coded genetic algorithm, *Applied soft computing*, 8, 2008, pp. 1105-1111.

[2] T.H. Cheng, Chih-Ping Wei, Vincent S. Tseng, Feature selection for medical data mining: comparisons expert judgment and automatic approaches, In: *Proceedings of the 19th IEEE symposium on computer-based medical system*, 2006, pp.165-170.

[3] R.E. Abdel-Aal, GMDH-based feature ranking and selection for improved classification of medical data, *Journal of Biomedical Informatics*, 38, 2005, pp.456-468.

[4] Z.Pawlak, Rough set, *International Journal of Computer and Information Sciences*, Vol.11, No.5, 1982, pp.341–356.

[5] Liu Yaohe, Wu Pei, Tan Baohua, The Application of medical data mining based on rough set. *Journal of HeBei University of technology*, Vol.23,No.2, 2008,pp.68-70.

[6] Qin Zhongguang, The crossing research on rough set and its application in traditional Chinese medicine diagnosis [D], *South China University of technology*, Guangzhou, China, 2002.

[7] S. Tsumoto, Mining diagnostic rules from clinical databases using rough sets and medical diagnostic model, *Information Sciences*, 162 ,2004, pp.65-80.

[8] Kim Soohwan, Jun soojin, Han Seonkwan, Rough set reasoning system for deciding learning style in cyber education, *WSEAS Transactions on Information Science and Applications*, Vol.4, No. 2, 2007, pp.324-330.

[9] J. Bazan, A comparison of dynamic and non-dynamic rough set methods for extracting laws from decision tables, in: L. Polkowski, A. Skowron (Eds.), Rough Sets in Knowledge Discovery, *Physica-Verlag, Heidelberg*, 1998, pp.321-365.

[10] Ming Yang, Ping Yang, A novel condensing tree structure for rough set feature selection, *Neurocomputing*,71,2008, pp.1092-1100.

[11] Juzhen Dong, Ning Zhong, Setsuo Ohsuga, Using rough sets with heuristics for feature selection, in: N.Zhong, A.Skowron, S.Ohsuga (Eds.), RSFDGrC'99, *Lecture Notes in Artificial Intelligence*, Vol. 1711, 1999, pp. 178-187.

[12] Zhao Rongyong, Zhang Hao, Li Cuiling, Lu Jianfeng, Wang Jun. Disjunctive Normal Form Generation Algorithm for Discernibility function in Rough Set Theory. *Computer Engineering*, Vol.32,No.2, 2006, pp.183-185.

[13] Chirnphlee Siriporn, Salim Naomie, Ngadiman Mohd Salihin Bin, et al., Rough fuzzy approach for web usage mining, *WSEAS Transaction on*

Information Science and Applications, Vol.3,No.3, 2006,pp.618-621.

[14] Zeng Zhiming, Jiang Ge, Disjunctive Normal Form Generation Algorithm in Rough Set Theory, *Fu Jian Computer*, 2,2008,pp.72,71.

[15] Zhang Wenxiu, Wu Weizhi, Liang Jieye, Li Deyu, Rough set Theory and Method, 2001, *Science Press*.

[16] Gao Kun, Chen Zhongwei, Liu Meiqun, Predicting performance of grid based on rough set, *WSEAS Transactions on Systems,* Vol. 7, No. 4, 2008, pp.288-297.

[17] Perez Rafael Bello, Nowe Ann, Vrancx Peter, et al., Using ACO and rough set theory to feature selection, *WSEAS Transaction on Information Science and Applications*, Vol.2, No.5, 2005,pp.512-517.

[18] Xiangyang Wang, Jie Yang, Richard Jensen, Xiaojun Liu, Rough set feature selection and rule induction for prediction of malignancy degree in brain glioma, *Computer methods and programs in biomedicine*, Vol.83,2006, pp.147-156.