IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

SPECIAL SECTION ON TRENDS, PERSPECTIVES AND PROSPECTS OF MACHINE LEARNING APPLIED TO BIOMEDICAL SYSTEMS IN INTERNET OF MEDICAL THINGS

Received July 10, 2018, accepted August 15, 2018, date of publication August 21, 2018, date of current version October 12, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2866389

Mining of Nutritional Ingredients in Food for Disease Analysis

ZHENFENG LEI[1], SHUANGYUAN YANG[2], HAN LIU[3], (Member, IEEE), SABA ASLAM[1], (Member, IEEE), JINYU LIU[4], HALEFOM TEKLE[1], EMMANUEL BUGINGO[1], AND DEFU ZHANG[1], (Member, IEEE)

[1]School of Information Science and Engineering, Xiamen University, Xiamen 361005, China
[2]Software School of Xiamen University, Xiamen University, Xiamen 361005, China
[3]School of Computer Science and Informatics, Cardiff University, Cardiff CF24 3AA, U.K.
[4]School of Engineering, Hong Kong University of Science and Technology, Hong Kong

Corresponding author: Shuangyuan Yang (yangshuangyuan@xmu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61672439 and in part by the Natural Science Foundation of Fujian Province, China, under Grant 2015J01288.

ABSTRACT Suitable nutritional diets have been widely recognized as important measures to prevent and control non-communicable diseases (NCDs). However, there is little research on nutritional ingredients in food now, which are beneficial to the rehabilitation of NCDs. In this paper, we profoundly analyzed the relationship between nutritional ingredients and diseases by using data mining methods. First, more than 7000 diseases were obtained, and we collected the recommended food and taboo food for each disease. Then, referring to the *China Food Nutrition*, we used noise intensity and information entropy to find out which nutritional ingredients can exert positive effects on diseases. Finally, we proposed an improved algorithm named CVNDA_Red based on rough sets to select the corresponding core ingredients from the positive nutritional ingredients. To the best of our knowledge, this is the first study to discuss the relationship between nutritional ingredients in food and diseases through data mining based on rough set theory in China. The experiments on real-life data show that our method based on data mining improves the performance compared with the traditional statistical approach, with the precision of 1.682. In addition, for some common diseases, such as diabetes, hypertension and heart disease, our work is able to identify correctly the first two or three nutritional ingredients in food that can benefit the rehabilitation of those diseases. These experimental results demonstrate the effectiveness of applying data mining in selecting of nutritional ingredients in food for disease analysis.

INDEX TERMS Attribute reduction, data mining, disease analysis, food, information entropy, nutritional ingredients, rough sets.

I. INTRODUCTION

NCDs are chronic diseases, which are mainly caused by occupational and environmental factors, lifestyles and behaviors, including Obesity, Diabetes, Hypertension, Tumors and other diseases. According to the *Global Status Report on Non-communicable Diseases* issued by the WHO, the annual death toll from NCDs keeps adding up, which has caused serious economic burden to the world. About 40 million people died from NCDs each year, which is equivalent to 70% of the global death toll [1]–[3]. Statistics of *Chinese Resident's Chronic Disease and Nutrition* shows that, the number of the patients suffering from NCDs in China is higher than the number in any other countries in the world, and the current prevalence rate has blown out. In addition, the population aged 60 or over in China has reached 230 million and about two-thirds of them are suffering from NCDs according to the official statistics [4], [5]. Therefore, relevant departments in each country, especially in China, such as medical colleges, hospitals and disease research centers all are concerned about NCDs.

Suitable nutritional diets play an important role in maintaining health and preventing the occurrence of NCDs [2], [6]. With the gradual recognition of this concept, China has also repositioned the impact of food on health. However, research on nutritional ingredients in food via data mining, which are conducive to the rehabilitation of diseases is still rare in China. At present, China has just begun the IT (Information Technology) construction of smart healthcare. Most studies on the relationship between nutritional ingredients in food and diseases are still through expensive precision instruments or long-term clinical trials. In addition, there are also many prevention reports, but they studied only

one or several diseases [7]. In China, studying the relationship between nutritional ingredients and diseases using data mining is immature. Most doctors only recommend the specific food to patients suffering from NCDs, without giving any relevant nutrition information, especially about nutritional ingredients in food [8].

The solutions for NCDs require interdisciplinary knowledge [2], [9]. In the era of big data, data mining has become an essential way of discovering new knowledge in various fields, especially in disease prediction and accurate health-care (AHC). It has become a core support for preventive medicine, basic medicine and clinical medicine research [10]. With respect to the disease analysis through the mining of nutritional ingredients in food, we mainly make the following contributions: (i) We extracted data related to Chinese diseases, corresponding recommended food and taboo food for each disease as many as possible from medical and official websites to create a valuable knowledge base that are available online; (ii) Applying noise-intensity and information entropy to find out which nutritional ingredients in food can exert positive effects to diseases; (iii) In this paper, the data is continuous and has no decision attributes. To address this problem, we proposed an improved algorithm named CVNDA_Red based on rough set theory, which can better select corresponding core ingredients from the positive nutritional ingredients in food.

The structure of this paper is organized as follows: Section II reviews the related work in the field of disease analysis and data mining. Section III describes the specific data mining algorithms used in this paper, reasons why we select the algorithms, as well as two evaluation indexes. Section IV elaborates the data, experimental results and analysis in detail. Section V presents discussions between methods. Some conclusions and potential future research directions are also discussed in Section VI.

II. RELATED WORK

With the continuous expansion of the impact of NCDs on people, various countries such as South Korea, the United States, Germany, Netherlands and China have conducted research on relationship between food and diseases in recent years. The United States used the *Framingham Cohort* to find out the relationship between the incidence of Coronary disease and nutritional diets [8], [11]. South Korea used factor analysis, reduced rank regression and the rating method to study the relationship between dietary patterns and nutrition-related health problems [12], [13]. Computational results in [14] showed, a daily intake of 100 grams of fresh fruits may reduce almost one third of the risk of Cardiovascular diseases. Li *et al.* [15] found that Vitamin D supplementation in Chinese population cannot completely improve the lack of Vitamin D. Agapito *et al.* [16] proposed that we can recommend food according to the body's Creatinine values. However, the above studies are basically carried out through long-term clinical trials, which just recommend food for certain specific diseases and they seldom study the relationship between nutritional ingredients and diseases by data mining techniques.

Large-scale and high-dimensional data that we extracted from online have usually a noise, hence we must use some techniques to pre-process the data. For example, the data denoising method based on wavelet transform [17] was proposed in the field of signal and data compression. Zhang and Zhang [18] proposed a new SVM-GARCH prediction model based on time series data to select features by mutual information. Diogo *et al.* [19] applied noise reduction to exchange rate chaotic data. Actually, the noise also can measure the fluctuations in gene expression, which is one of the important biology indicators. Therefore, many literatures indicate that they are currently utilizing the noise-intensity to discover new knowledge [20]. For example, the concrete expression of protein noise and noise-intensity at equilibrium were described [21], [22], and they also analyzed kinetic behavior by the detailed expression at non-equilibrium. Lei *et al.* [23] used noise-intensity to improve the LDA algorithm to achieve better performance on a protein classification task according to the protein distribution structure.

Information entropy is a measurement to eliminate uncertainty from given information. Since the concept of entropy was introduced into information theory, it has been widely used in different fields. Yang and Jiang [24] applied the maximum fuzzy entropy and fuzzy c-means segmentation [25] to select the thresholds automatically. An image classification method of invariant pixel region texture based on wavelet packet entropy was also proposed [26]. Bakhshali [27] combined the MFE criterion with the MMI criterion to present a maximum fuzzy segmentation algorithm based on mutual information. A method of recovering short vowels and other phonetic symbols from Arabic phonetic symbol documents was proposed by using the maximum entropy model [28]. Tan and Taniar [29] proposed a distribution estimation algorithm based on maximum entropy, and they demonstrated experimentally the superior performance in dealing with clustering problems. An optimization strategy based on maximum entropy in [30] was proposed to process the parallel resource. Chen [31] described a MAODV protocol based on entropy in Ad Hoc networks and proposed a new measure to express the path stability. An extraction method in [32] was presented for pruning larger nodes with information entropy. John *et al.* [33] used the utility value reflected by information entropy to calculate the weight of evaluation indexes and proposed an AHP fuzzy comprehensive evaluation model. This model can provide a flexible tool for analysts to strengthen security of the port systems.

In rough set theory, the attribute reduction of decision tables is an interesting research topic. In general, we used heuristic methods to reduce attributes. Firstly, the core attributes must be identified, from these the most important attributes must be selected, and finally added to the reduction set [34]. For example, a method based on discernibility matrix and logic operation was given [35]. Nguyen [36] proposed two fusion methods based on logical

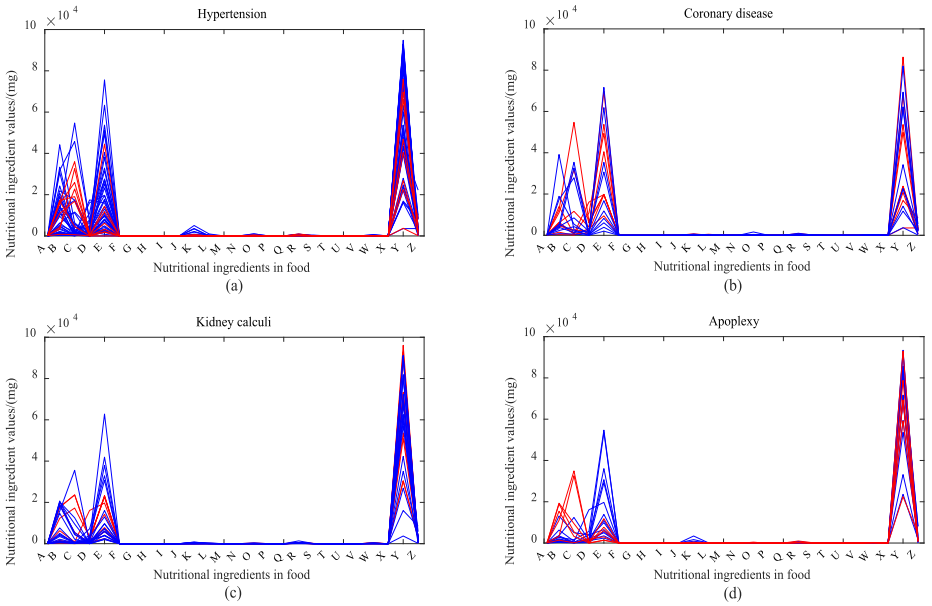


FIGURE 1. The distribution of nutritional ingredient values in recommended food and taboo food.

reasoning and discernibility matrix. An inductive learning method based on rough sets was proposed for SAR analysis [37]. Zhang *et al.* [38] proposed a multivariate decision tree method by using rough set theory. An improved algorithm for finding the minimum attribute reduction set was presented by Jia *et al.* [39] using the decision table's generalized information. In addition, Zheng [40] defined the attribute significance from the perspective of mutual information, and proposed a MIBK algorithm to reduce attributes. The idea of functional decomposition in [41] and [42] was applied to the decomposition of decision tables. They established a hierarchical decision model by gradually delaminating from an inaccurate single large-scale objective into sub-goals, which improves the efficiency of data analysis. However, most of the above methods are basically based on the decision tables, which are discrete and have decision attributes. These methods will not work well if there is a decision table with continuous values and non-decision attributes. To solve this problem, we proposed an algorithm named CVNDA_Red based on rough sets, which is suitable for continuous values as well as non-decision attributes and can better select corresponding core nutritional ingredients in food for diseases.

III. METHODOLOGY

To show whether data mining methods can be applied to the disease analysis or not, we adopted different methods for research purpose. Since this paper is trying to solve a new problem from a real-world application, there is no relevant work for comparison. But we used an exploring way to solve this problem, in other words, we conducted an effect comparison between different methods to select the best one. In order to show the characteristics of nutritional ingredients in food, we draw nutritional ingredient values in recommended food and taboo food for four common diseases. In Fig.1, it shows the nutritional ingredient values in the recommended food and taboo food for four diseases, such as Hypertension, Coronary disease, Kidney calculi and Apoplexy. The X-axis and Y-axis represent nutritional ingredients and corresponding values respectively. The meaning of each letter of X-axis is expressed in Table 2. In addition, blue and red represent recommended food, taboo food respectively. It is clear from Fig.1 that some nutritional ingredient values are very high and the others are very low, no matter in recommended food or taboo food. Generally, there are few taboo food for most diseases in China, as a result we can not collect more. Given the less credible knowledge for small sample sizes by data mining, we are currently focusing our research on recommended food to find which nutritional ingredients can exert positive effects for a certain disease. In this paper, the nutritional ingredients that benefit the rehabilitation of diseases are called positive nutritional ingredients (PNIs). On the contrary, the nutritional ingredients that aggravate diseases are named negative nutritional ingredients (NNIs).

A. STATISTICAL ALGORITHM (SA)

Fig.1 clearly shows that nutritional ingredient values vary widely. If a certain disease is caused by the lack of certain nutritional ingredients, then their values in recommended food should be relatively higher in theory. Therefore, we can figure out which nutritional ingredient values are high. The nutritional ingredients with higher values should be the ingredients, which benefit the rehabilitation of that specific disease, i.e., PNIs [43]. This method is referred to as statistical algorithm (SA) in this paper. The common notations are defined as following, let n be the number of the recommended food for a certain disease, and m be the number of nutritional

ingredients. The above idea can be expressed as below:

$$sort\{\sum_{i=1}^{n} x_{i1}, \sum_{i=1}^{n} x_{i2}, \cdots, \sum_{i=1}^{n} x_{im} \mid descend\} \quad (1)$$

where x_{i1} denotes the first nutritional ingredient value of the i^{th} recommended food for a certain disease; x_{i2} indicates the second nutritional ingredient value of the i^{th} recommended food, and so on. All recommended food for this disease is summed according to nutritional ingredient values. Then, the m summations are sorted in descending order. So, from the sorted nutritional ingredients the most top ones should be PNIs by this way.

B. NOISE-INTENSITY ALGORITHM (NI_SA)

If a certain disease is caused by the lack of some nutritional ingredients, the recommended food should contain these ingredients. In other words, their certain values may not be the highest, but must be existing in all recommended food (high stability), otherwise these recommended food have no reason to be recommended. Conversely, if these nutritional ingredients are not PNIs for that specified disease, they may or may not exist in different recommended food (poor stability). The SA just considers the level of nutritional ingredient values to determine simply whether they are PNIs or not. Therefore, on the basis of SA, we can further consider about the stability of nutritional ingredient values in food for determining whether they are PNIs or not. In this way, two aspects can be considered, i.e., the level and stability of nutritional ingredient values. The greatest advantage of using this method is that it reduces the error rate of what are considered to be PNIs just because of the level of nutritional ingredient values.

In biology, noise-intensity is a measure of stability. As a matter of fact, the noise-intensity can be considered as reflecting the degree of data dispersion, because it is not only affected by the variance, but also controlled by the mean [44]. Therefore, the purpose of using noise-intensity in this method is to express the stability of nutritional ingredient values. It is worth noting that the noise-intensity calculated here is based on the original data, which is raw data, because the variance of processed data is one. The noise-intensity and above idea are expressed as follows:

$$\varphi_j = \frac{\delta_j}{\mu_j} (j = 1, 2, 3, \cdots, m) \quad (2)$$

$$sort\{\varphi_1 \cdot \sum_{i=1}^{n} x_{i1}, \varphi_2 \cdot \sum_{i=1}^{n} x_{i2}, \cdots, \varphi_m \cdot \sum_{i=1}^{n} x_{im} | descend\} \quad (3)$$

where φ_j represents noise-intensity of the j^{th} nutritional ingredient; δ_j indicates the variance of the j^{th} nutritional ingredient and μ_j is the corresponding mean. The noise-intensity takes the stability of each nutritional ingredient into account. In this paper, each noise-intensity is treated as a weight scale factor for each corresponding nutritional ingredient summation. This method based on SA is called noise-intensity algorithm (NI_SA).

C. INFORMATION ENTROPY ALGORITHM (IE_SA)

Fig.1 clearly illustrates that there are two or more nutritional ingredients with higher values in a disease. For example, the Y nutritional ingredient and the E nutritional ingredient are relatively higher than other nutritional ingredients in Hypertension. Similarly, the stability of nutritional ingredients follows the same statistics. To find out which one is a PNI in more than two nutritional ingredients. Based on their similarity value levels or stability, the validity of each nutritional ingredient is considered. For example, if a nutritional ingredient has a good effect to a certain disease, the corresponding validity should be greater than others.

Information entropy measures the degree of disorder of data systems and calculates the amount of valid information. The smaller the information entropy is, the greater validity it provides and vice versa [45]. In this paper, information entropy calculates the validity of each nutritional ingredient. According to the definition of entropy, the entropy of the nutritional ingredient can be expressed as follows:

$$H_j = -\frac{1}{\ln n} \sum_{i=1}^{n} f_{ji} \ln f_{ji} \quad (4)$$

In the definition of entropy, f_{ji} should be considered as the occurrence probability of the j^{th}nutritional ingredient of the i^{th} food. Since our data does not have identical food in a disease (i.e., the occurrence probability of each food in a disease is one), Equation (4) is not acceptable in our paper. However, the above mentioned occurrence probability can be replaced by each nutritional ingredient value. Because it makes sense that if there is a higher nutritional ingredient value in food, the occurrence probability is larger. In order to make $\ln f_{ji}$ reasonable, if $f_{ji} = 0$, we would note $f_{ji} \ln f_{ji} = 0$. Then, f_{ji} is redefined as follows:

$$f_{ji} = \frac{x_{ji}}{\sum_{i=1}^{n} x_{ji}} \quad (5)$$

where x_{ji} represents the value of the j^{th} nutritional ingredient of the i^{th} food. The validity of the j^{th} nutritional ingredient can be defined as follows:

$$W_j = \frac{1 - H_j}{m - \sum_{j=1}^{m} H_j} \quad (6)$$

$$sort\{W_1 \cdot \sum_{i=1}^{n} x_{i1}, W_2 \cdot \sum_{i=1}^{n} x_{i2}, \cdots, W_m \cdot \sum_{i=1}^{n} x_{im} | descend\} \quad (7)$$

In this paper, we represent the validity as a weighted scale factor for each corresponding nutritional ingredient summation. This method based on SA is called information entropy statistical algorithm (IE_SA).

D. NOISE-INTENSITY AND INFORMATION ENTROPY (NIIE_SA)

As mentioned above, NI_SA and IE_SA are both based on the SA method, but their starting points are different. One consideration is the stability after the level of the nutritional ingredient values, and the other one is to consider the validity

after the level of the nutritional ingredient values. For integrating the above two methods through a certain way such as voting, the results of each method can be considered as a kind of voting results. If both methods assume a certain nutritional ingredient having positive effects, it is more likely to consider that nutritional ingredient as a positive nutritional ingredient. It is more reasonable than deciding which one of those nutritional ingredients is a PNI by using a single method [46]. The above idea can be expressed as follows:

$$sort\{\begin{bmatrix} sort_{11} \\ + \\ sort_{21} \end{bmatrix}, \begin{bmatrix} sort_{12} \\ + \\ sort_{22} \end{bmatrix}, \cdots, \begin{bmatrix} sort_{1m} \\ + \\ sort_{2m} \end{bmatrix} | ascend\} \quad (8)$$

where $sort_{11}$ represents the ranked serial number of the first nutritional ingredient according to NI_SA; $sort_{21}$ shows the ranked serial number of the first nutritional ingredient according to IE_SA, and so on. The sorting strategy is as follows: summing up their ranked serial numbers for the same nutritional ingredients, and then sorting the gained summations in ascending order. According to the above two methods, it is clear that the greater possibility of nutritional ingredients are viewed as positive nutritional ingredients, the smaller summations are sorted, i.e., the top ranked ingredients should be PNIs. In this paper the method based on NI_SA and IE_SA is referred to as NIIE_SA.

E. ROUGH SET ALGORITHM (RS)

As can be seen from Table 3, the number of positive nutritional ingredients is different for each disease. For some diseases, 10 out of 26 nutritional ingredients (almost $\frac{2}{5}$) are PNIs, while others have only two PNIs. So, can some nutritional ingredients with little positive effects be excluded first? In this way, the rest nutritional ingredients are almost positive to diseases, and the same is achieved for our purposes. In this paper, this method based on attribute reduction is called RS.

The rough set theory is a mathematical tool developed by a Polish computer scientist, Pawlak [47], to study uncertain and inaccurate knowledge [48]. Researchers have proposed lots of attribute reduction algorithms, such as based on positive domain, discernibility matrix, and decision tree. The definition of rough sets is as follows. The form of a decision table in rough sets can be represented by a four-tuples: $S = (U, A, V, f)$, which is also expressed as $S = (U, V)$. U called the universal set represents a non-empty, finite set of objects. A denotes a non-empty, finite set of attributes, $A = C \cup D$, where C is a set of condition attributes, and D is a set of decision attributes. If $D = \phi$, then, S represents a decision table without decision attributes. $V = \bigcup_{a \in A} V_a$ is a set of attribute values, V_a denotes values that the attribute a may take. $f : U \times A \to V$ is an information function that assigns a value to each attribute of each object, i.e., $\forall a \in A, x \in U, f(x, a) \in V_a$.

Let P be a subset of A ($P \subseteq A$), the associated equivalence relation $ind(P)$ can be expressed as: $ind(P) = \{(x, y) \in U \times U \,|\, \forall a \in P, f(x, a) = f(y, a)\}$. $ind(P)$ is called a $P-$indiscernibility relation on U. $U/ind(P)\,(or\ U/P)$ represents a set, which uses $ind(P)$ to classify U. The equivalence classes of the P-indiscernibility relation are denoted as $[x]_P = \{y \,|\, y \in U, (x, y) \in ind(P)\}$. Dependency of the condition attribute P in the decision attribute D can be expressed as $|POS_P(D)|/|U|$. It represents the positive domain of P in D. In these sets, P can uniquely classify all elements in U as a subset U/D. The significance degree $sig(a)$ of the attribute a in the attribute set P can be defined as follows:

$$sig(a) = |POS_P(D) - POS_{P-a}(D)|/|U| \quad (9)$$

$$tol(S) = |POS_C P(D)|/|U| \quad (10)$$

where $tol(S)$ is called the compatibility degree of S, if $|POS_C(D)| = |U|$, the decision table is considered to be a compatible one, otherwise it is incompatible. Obviously, the necessary and sufficient condition of a compatible decision table is that its compatibility degree equals to one. For all decision tables, contribution of each condition attribute on the decision attributes is different. The contributions of condition attributes to decision attributes are called the attribute significance. The attribute reduction algorithms are reflected by the changes of the classification ability after the deletion of a certain attribute. However, the above attribute reduction algorithms are based on decision tables whose attribute values are discrete and which have decision attributes. The data in this paper is obviously not suitable for those methods. At present, attribute reduction methods with continuous values are through discretization processing, and then to select features without supervision. But the results of discretization are bound to lose a lot of information in this way. So, direct reduction of attributes in the case of continuous values will be more realistic. To this end, we presented an improved algorithm (i.e., called continuous value and no decision attributes reduction, CVNDA_Red) to suit this kind of data. Firstly, establishing food fuzzy similar matrix is needed [49]:

$$Sim(o_s, o_t) = \min\{1 - \frac{|x_{sj} - x_{tj}|}{\max x_j - \min x_j}\} \quad (11)$$

where $Sim(o_s, o_t)$ denotes the similarity of food s and t. The $\max x_j$ and $\min x_j$ represent the maximum value and the minimum value of the j^{th} nutritional ingredient in food for a disease respectively. Then, the square method is used to find the transitive closures of the fuzzy similarity matrix. The aim is to extract the different values of λ from transitive closures. After sorting them in descending order, it needs to calculate the corresponding compatibility degree based on λ. This step will be terminated until it is found that the compatibility degree of decision tables is less than one. At this time, λ can be viewed as a next step threshold. In this paper, since there is no decision attributes in data, we treated all condition attributes as decision attributes.

For each condition attribute $a\,(a \in C)$, after deleting a, the calculation of the corresponding similarity matrix and transitive closures is needed by the same ways as above in the case of the obtained threshold, and then to cluster the results $(C - \{a\})$. Finally, the significance degree of condition attribute a is calculated, and other attributes are followed

by analogy. The definition of the significance degree shows that, if the deletion of a certain condition attribute has a higher influence on the classification performance under the same threshold, the significance of this attribute is higher. Otherwise, it is smaller. Those attributes whose significance degrees all equal to zero are deleted to achieve the purpose of attribute reduction. The specific steps of this algorithm are described as follows:

Algorithm 1: CVNDA_Red

Input: Food-nutritional ingredients matrix ($n \times m$)
Output: sn % the serial numbers of selected nutritional ingredients

1 $R \leftarrow Sim(Data)$; % R is a symmetric matrix ($n \times n$)
2 $S \leftarrow transClosure(R)$;% obtaining transfer closure matrix by square method
3 $\lambda \leftarrow sort(unique(S), 'descend')$;
4 **for** $i = 1 : size(\lambda)$ **do**
5 $\quad xrd = tol(\lambda_i)$;
6 $\quad$ **if** $xrd < 1$ **then**
7 $\quad\quad yz = \lambda_i$; % getting the threshold
8 $\quad\quad break$;
9 $\quad$ **end**
10 **end**
11 $SIG = zeros(1, m)$;
12 **for** $i = 1 : m$ **do**
13 $\quad R \leftarrow Sim(Data - a)$; % obtaining fuzzy similar matrix after deleting attribute a
14 $\quad S \leftarrow transClosure(R)$;
15 $\quad \lambda \leftarrow sort(unique(S), 'descend')$;
16 $\quad$ **for** $i = 1 : size(\lambda)$ **do**
17 $\quad\quad$ **if** $\lambda_i > yz$ **then**
18 $\quad\quad\quad$ % nothing need to do!
19 $\quad\quad$ **else**
20 $\quad\quad\quad U = union(U, cluster(\lambda_i))$;
21 $\quad\quad$ **end**
22 $\quad$ **end**
23 $\quad SIG(1, i) = sig(U)$;
24 **end**
25 $sn = greatZero(SIG)$;
26 **return** sn

In Algorithm 1, Line 3 calculates different values of λ according to the transitive closure matrix obtained from Line 2; Line 4-10 obtain the threshold by compatibility degree of decision tables (less than one); Line 16-22 get the equivalence classes according to the λ_i, and then cluster the results. The last clustering results are unionized with the next clustering results until λ_i is greater than the threshold yz; Line 25 chooses the serial numbers of nutritional ingredients whose significance degrees are greater than zero.

F. COMBINATION OF NIIE_SA AND ROUGH SETS (RSNIIE_SA)

It is clear from RS that the obtained results are only nutritional ingredients, which are viewed as positive nutritional ingredients, but they are not sorted according to the possibility of judging as PNIs. So, using these results directly on patients will not be effective. In order to more fully identify nutritional ingredients, which have the most effects on the rehabilitation of diseases, results obtained by RS can be sorted. The sorting rule is based on the ranked serial numbers of nutritional ingredients of NIIE_SA. For example, if the nutritional ingredients obtained by RS for Hypertension are A, B and C nutritional ingredients, and the ranked serial numbers of those ingredients obtained by NIIE_SA are the third, the second, and the first respectively, then, the results obtained using this method are C, B and A. In this paper, this method based on RS and NIIE_SA is called RSNIIE_SA.

Fig.2 gives better understanding of the relationship among the above methods intuitively. According to the definition of RS, the results are not sorted according to the possibility of judging as PNIs. Therefore, the results of RS are not presented as the final results here.

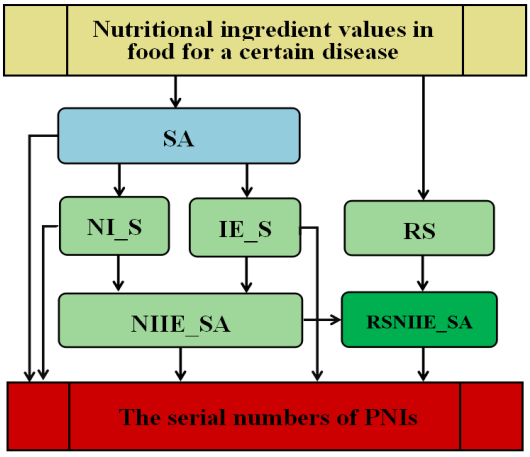


FIGURE 2. The relationship among methods used in this paper.

IV. EXPERIMENTS

A. DATA SET

We mainly used the combination of Spider technology [50] and Selenium 2 technology [51] based on the Python language to get the data. The main form to get Spider data is the texts that we extracted from web pages. Firstly, we downloads the relevant web pages according to specified URLs, and then uses regular expressions to locate the desired texts. But nowadays it is not easy to get data using Spider as many websites deploy anti-crawling technology. However, another technology called Selenium 2 can be used to extract the data as it is automatic to simulate the people's behaviors to access to websites. We can also get data from academic papers or articles. In order to obtain high quality data, we have mainly chosen Chinese medical and official websites, including:

- 360baike (*https://baike.so.com*);
- CDC (*http://www.chinacdc.cn*);
- XYWY (*http://jib.xywy.com*);

TABLE 1. The instances of data used in this paper.

Disease	Recommended food	Taboo food
Hypertension	Reishi, Agaric, Mushroom, Celery, Leek, Cole rape, Day lily, Shepherd'purse, Chinese cabbage, Crown daisy, Amaranthus tricolor ...	Chocolate, Wine, Tea, Pickle, Salted fish, Bacon, Ham, Marinated duck, Dog meat ...
Coronary disease	Kelp, Laver, Agaric, Sesame, Walnut, Fish oil, Yam, Haw flakes, Garlic, Corn, ...	Crab, Oyster, Squid, Speck, Yolk, Cream, Tea, Candy, Wine ...
Kidney stone	Broccoli, Prunus armeniaca, Cantaloupe, Agaric, Pumpkin, Chicken liver, Watermelon, Carrot ...	Beef, Mutton, Wine, Onions, Garlic, Pickle, Bacon, Fish ...
.....	.....	.....

- 360liangyi (*http://ly.so.com*);
- baiduxueshu (*http://xueshu.baidu.com*);
- baikemingyi (*http://www.baikemy.com*);
- and so on.

All these websites are available online. We follow two steps to obtain the data. Firstly, all the possible diseases were collected, and then the recommended and taboo food were fetched according to the corresponding diseases. We have collected about 12,100 diseases, out of them 9,100 diseases were selected by removing repeated diseases. Based on 9,100 diseases, we have also collected 56,000 data that contains a number of recommended and taboo food. While selecting recommended food or taboo food for each disease, we adopted the principle of intersection to deal with the data, i.e., it is selected if a certain food showed up in each piece of data for a disease. Due to the existence of diseases without recommended and taboo food, the total instances of data become 7,594 after we remove these diseases. Each one of these data includes three items: the disease name, recommended food, and taboo food. Here are several examples:

The referred nutritional ingredient values for each food are from the *China Food Composition 2002* [52] and the *China Food Composition 2004* [53], which are compiled by the *National Institute for Nutrition and Health Chinese Center for Disease Control and Nutrition and Food Safety*. The former version includes sixteen nutritional ingredients, whereas the later version contains seventeen nutritional ingredients. The common data we got in the two versions have been deleted and we summed up the remaining data that results 26 nutritional ingredients in food, which finally are presented as follows:

The capital letters represent 26 nutritional ingredients. Since the representation of Vitamin A and Retinol equivalence in China is not consistent and they have different values for some certain food in the two versions, two different nutritional ingredients were considered. In order to avoid missing any possibility, we assumed them as two PNIs if any one of them has positive effects for diseases. If nutritional ingredient values of a certain food are not available in the two versions, a weighted average value of similar food will be filled, otherwise that food will be deleted.

TABLE 2. The description (letter representation, name, unit) of 26 nutritional ingredients in food.

Label	Nutritional ingredient	Unit	Label	Nutritional ingredient	Unit
A	Energy	kcal	N	Vitamin C	mg
B	Protein	g	O	Cholesterol	mg
C	Fat	g	P	Renieratene	μg
D	Fiber	g	Q	Retinol equivalent	μg
E	Carbohydrate	g	R	Potassium	mg
F	Vitamin A	μg	S	Magnesium	mg
G	Vitamin B1	mg	T	Manganese	mg
H	Vitamin B2	mg	U	Zinc	mg
I	Niacin	mg	V	Copper	mg
J	Vitamin E	mg	W	Phosphorus	mg
K	Sodium	mg	X	Selenium	μg
L	Calcium	mg	Y	Water	g
M	Iron	mg	Z	Ash specification	g

In order to verify the rationality of the above methods, ten common diseases are selected as testing samples as shown in Table 3. These diseases were selected because of their unquestionable PNIs and NNIs, either in China or abroad. Of course, we also proved the correctness of the testing samples below by the relevant medical experts. Since there are no unified conclusions about other diseases, those cannot be verified yet. But for future work, many other examples can be taken to implement the above methods and to verify the feasibility. In this paper, the first letter of the word *Disease* as well as the word *Name* and serial number are used to indicate the corresponding diseases.

Vitamin is a set that includes F, G, H, J, N nutritional ingredients; Vitamin B contains G, H nutritional ingredients. Since there are not statistics about Vitamin D and Mineral salt nutritional ingredient values in food, this paper does not discuss whether these two nutritional ingredients are helpful to rehabilitate the diseases or not.

B. EVALUATION MEASURE

In order to explain and compare the validity of data mining methods applied in our study, two evaluation indexes, individual precision (*IP*) and over precision (*OP*), are proposed. They are defined as follows, which are based on the idea that the top ranking nutritional ingredients sorted by a certain method have the greater possibility of being judged as PNIs. In order to do the mathematical calculation, each value of positive nutritional ingredients is noted as *one* from the obtained results (i.e., the blue color in the experimental results below), and each of the others is noted as *zero*. For RS and RSNIIE_SA algorithm, the results of each disease may only

TABLE 3. The positive nutritional ingredients and negative nutritional ingredients for ten diseases.

Label	Disease	PNIs	NNIs
DN1	Nearsightedness	Calcium, Vitamin B, Vitamin A, Vitamin C, Vitamin E, Vitamin D	Carbohydrate
DN2	Caries	Magnesium, Phosphorus, Vitamin	Carbohydrate
DN3	Chronic cold	Vitamin, Zinc, Selenium	Sodium, Fat, Carbohydrate
DN4	Hypertension	Fiber, Potassium, Magnesium, Calcium	Sodium, Carbohydrate
DN5	Acne	Vitamin B, Vitamin A, Zinc	Fat, Energy, Carbohydrate
DN6	Coronary disease	Vitamin C, Calcium	Cholesterol, Fat
DN7	Kidney stone	Vitamin, Magnesium	Sodium, Calcium
DN8	Diabetes	Fiber, Protein, Vitamin, Mineral salt	Fat, Energy
DN9	Apoplexia	Potassium, Vitamin C, Magnesium	Cholesterol, Sodium
DN10	Anemia	Iron, Copper, Zinc, Protein, Vitamin B, Renieratene	Calcium, Phosphorus

select a few nutritional ingredients, and we would fill with zero here if other nutritional ingredients had no values (*zero* or *one*). Because if a certain nutritional ingredient was not selected, it would mean that the possibility of being assigned as a PNI is *zero*.

$$IP_k = \sum_{j=1}^{m} B_{kj}(\frac{1}{j})^l \tag{12}$$

where IP_k represents the individual precision of the k^{th} disease, B_{kj} denotes a binary value (*one* or *zero*) of the j^{th} nutritional ingredient. l represents the value coefficient ($l \geq 1$), which is added to emphasize the high value of the top ranking nutritional ingredients in experimental results [54], [55]. The value of l in this paper is two. Of course, in order to emphasize the higher value, it could be made larger. According to actual needs, the value of l can be changed. The IP index is mainly to compare the ability of different methods to find out PNIs for a certain disease. According to the Equation (12), the increasing rate will be very slow if the first value is zero, especially when l is larger (i.e., the purpose is to emphasize the significance of accurately choosing nutritional ingredients as PNIs in the first several times.[1]) It is the same in real cases. For example, when a doctor is recommending some food to a patient, he must first choose the food that contains rich PNIs to that patient. If the doctor's first recommended food does not satisfy patients' condition, the patients will definitely blame that the doctor did not identify the real cause of disease. Therefore, the above proposed index IP makes sense.

$$OP = \sum_{j=1}^{m}[all \cdot \sum_{k=1}^{g} B_{kj}(1 - \frac{m'_k}{m})](\frac{1}{j})^l \tag{13}$$

where $all = 1/\sum_{k=1}^{g}(1 - \frac{m'_k}{m})$ represents the reciprocal of proportion summation of positive nutritional ingredients. g denotes the number of testing samples. m'_k indicates the PNIs number of the k^{th} disease. The OP is to compare the ability of different methods to find out positive nutritional ingredients for all diseases. According to Table 3, it can be seen that the number of PNIs for each disease is not equal. It is obviously unreasonable if all selected nutritional ingredients are viewed as *one*. For example, there are twenty positive nutritional ingredients for one disease and only two PNIs for other diseases. If there is only one chance to select a positive nutritional ingredient, obviously the correct rate of the disease with twenty positive nutritional ingredients is bigger than other diseases that have only two PNIs. Thus, *all* coefficient is added for avoiding this error.

[1]The times mentioned here and below refer to the number of columns displayed in experimental results. For example, in Fig.3, the results of SA method at the first time are all the Y nutritional ingredient for each disease, i.e., the first column.

C. EXPERIMENTAL RESULTS

1) THE RESULTS OF SA

Nutritional ingredients with blue color indicate PNIs (see Table 3, Column three). The expected experimental result shows that the nutritional ingredients on the left side of each row should be positive nutritional ingredients, i.e., the blue ones are transferred to the left. It is clear from Fig.3 that these are not the expected results. Instead, each row of the left side is filled with Y, E and B nutritional ingredients rather than PNIs. However, those results are consistent with Fig.1, in which the highest nutritional ingredient values for each disease is Y. It is not difficult to explain why the ranking results come out like that, i.e. nutritional ingredients with higher values are all Y, E and B in four diseases.

DN1	Y	E	B	C	D	Z	R	A	K	W	L	S	O	N	M	J	I	U	T	V	H	Q	F	P	G	X
DN2	Y	E	B	Z	C	D	K	R	L	W	A	O	S	N	M	J	I	U	P	T	V	Q	H	F	G	X
DN3	Y	E	B	C	D	Z	R	A	K	L	W	S	N	M	J	O	U	I	P	T	V	H	F	Q	G	X
DN4	Y	E	B	C	Z	D	K	R	A	W	L	O	S	N	M	J	I	U	T	P	V	H	Q	F	G	X
DN5	Y	E	C	B	D	Z	R	A	K	L	W	S	M	J	N	U	T	I	O	V	H	P	G	F	Q	X
DN6	Y	E	C	B	D	Z	R	A	K	L	W	S	M	J	N	U	T	I	O	V	H	P	G	F	Q	X
DN7	Y	E	B	C	D	Z	R	A	W	K	L	S	O	N	M	J	I	U	V	P	T	F	Q	H	G	X
DN8	Y	E	B	D	C	Z	R	A	W	S	L	K	N	M	J	U	T	I	V	O	G	H	P	Q	F	X
DN9	Y	E	D	B	Z	C	K	R	A	L	W	S	N	M	J	I	U	T	P	V	O	H	G	Q	F	X
DN10	Y	B	E	C	D	Z	R	O	W	K	A	L	S	M	I	J	U	F	Q	T	V	H	G	N	P	X

FIGURE 3. The ranking results based on SA.

At present, people are not free to claim only a few nutritional ingredients in food (e.g., I just want to eat the Vitamins inside of fruits or vegetables.). Only professional doctors can do that, but the price is too high. Now, doctors may

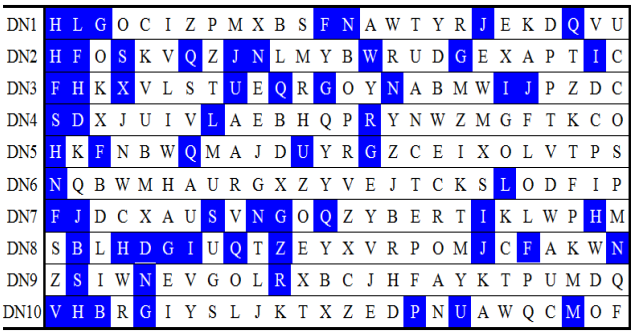


FIGURE 4. The ranking results based on NI_SA.

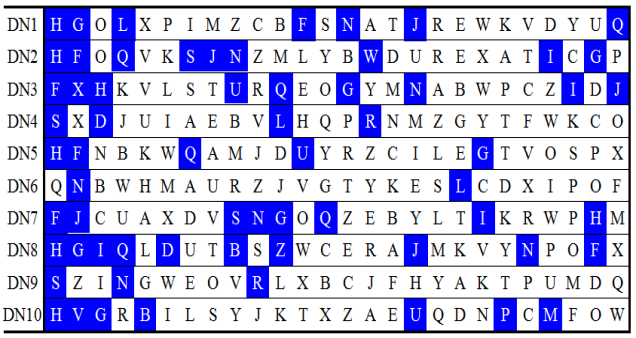


FIGURE 5. The ranking results based on IE_SA.

only consider to recommend these food with rich nutritional ingredients what you may lack, regardless of other nutritional ingredients whether you may lack or not. Therefore, SA is not feasible for selecting which ones are PNIs among many nutritional ingredients in first several times. In other words, only relying on statistical methods can't solve our problem.

2) THE RESULTS OF NI_SA

Fig.4 clearly shows that considering the stability of nutritional ingredients in food has great effects. Except for DN8 and DN9, the top ranking nutritional ingredients of each disease have become positive nutritional ingredients in the first time, and even the second ranking nutritional ingredients also have turned into what are expected, i.e., PNIs. Especially for DN1 and DN10, three out of seven PNIs (almost 50%), are selected successfully in first three times. Then, the ranked positions of PNIs are basically shifted left, which indicates that considering stability is a more reasonable direction.

3) THE RESULTS OF IE_SA

Fig.5 clearly shows that considering the validity of nutritional ingredients in food also leads to better results. Every first position of each row is placed with a positive nutritional ingredient except for DN6. And the second position of each row is also correct except DN4 and DN9. For DN8, the first four consecutive nutritional ingredients are all selected successfully in the first four times. Overall, this method is better than the above two methods (SA and IE_SA), which indicates

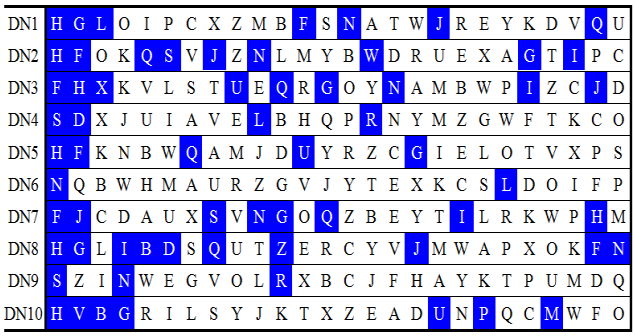


FIGURE 6. The ranking results based on NIIE_SA.

TABLE 4. The correct ratio for each disease.

DN1	DN2	DN3	DN4	DN5
$\frac{5}{9}$	$\frac{6}{12}$	$\frac{6}{13}$	$\frac{2}{5}$	$\frac{3}{6}$
DN6	**DN7**	**DN8**	**DN9**	**DN10**
$\frac{1}{4}$	$\frac{5}{10}$	$\frac{8}{13}$	$\frac{2}{4}$	$\frac{4}{9}$

that considering the validity of nutritional ingredients in food is still a relatively reasonable direction.

4) THE RESULTS OF NIIE_SA

Fig.6 elaborates that the use of voting strategy has better effects in determining which nutritional ingredients are most likely to become positive nutritional ingredients. This strategy turns the first ranked position of all diseases into a PNI, and the overall PNIs have moved to the left. This ensures that those PNIs can be identified correctly in the first time, so it is feasible to apply this strategy to solve our problem.

5) THE RESULTS OF RS

According to the RS algorithm, in each row of the above results, there is no significance difference among the selected nutritional ingredients. As can be seen from Fig.7, if only the nutritional ingredients are showing up in the experimental results, they are all considered as PNIs. It can be seen that most selected nutritional ingredients are correct. In order to explain the ability of this method in detail, we used the correct ratio (PNIs / selected nutritional ingredients) for each disease to show as follows:

The bold numbers indicate that more than half of the selected nutritional ingredients are correct. It clearly shows that the correct ratio of six diseases is greater than or equal to 50% and three other diseases (italic numbers) are close to 50%. So it can be demonstrated that most selected nutritional ingredients are correct, and this leads to the success of this method in finding out positive nutritional ingredients for diseases.

6) THE RESULTS OF RSNIIE_SA

The final and satisfactory results are shown in Fig.8. The first column shows that the results obtained in the first time, that

DN1	C	G	Y	P	H	F	E	L	Q				
DN2	H	F	M	Z	S	V	J	O	Q	Y	B	W	
DN3	F	H	X	R	Q	G	U	E	K	V	T	O	Y
DN4	X	D	S	H	G								
DN5	Q	A	K	F	H	C							
DN6	U	Q	N	K									
DN7	F	J	B	U	N	G	O	Q	C	E			
DN8	R	Z	B	D	S	Q	U	I	H	C	F	N	K
DN9	S	W	N	B									
DN10	H	D	U	N	S	A	B	L	V				

FIGURE 7. The experimental results based on RS.

DN1	H	G	L	P	C	F	E	Y	Q				
DN2	H	F	O	Q	S	V	J	Z	M	Y	B	W	
DN3	F	H	X	K	V	T	U	E	Q	R	G	O	Y
DN4	S	D	X	H	G								
DN5	H	F	K	Q	A	C							
DN6	N	Q	U	K									
DN7	F	J	C	U	N	G	O	Q	B	E			
DN8	H	I	B	D	S	Q	U	Z	R	C	K	F	N
DN9	S	N	W	B									
DN10	H	V	B	L	S	A	D	U	N				

FIGURE 8. The ranking results based on RSNIIE_SA.

means the PNIs for each disease. The results we obtained for the second time are shown in column two, but the results in the DN6 disease are not expected. Generally, this method gives the most accurate results among all the other methods, because we can accurately find out the first two or three nutritional ingredients out of 26 ingredients for each disease. It is the best method that meets our actual needs. For example, if the patient is suffering from the above ten diseases, the recommended food must be rich with the first two or three nutritional ingredients.

V. DISCUSSION

The above results are given independently, and no comparison is made among the methods. So, it is difficult to interpret which method is the best to select positive nutritional ingredients for diseases. According to the above two indexes, the next is a comparison of five methods based on the effect. The results we got from RS could not be directly used, as a result we are not able to plot as graph.

The graph in Fig.9 above the *reference line* (dotted line) indicates that the positive nutritional ingredients are selected successfully in the first time and the graph below the *reference line* represents PNIs, which are not selected successfully in the first time. The points near the *reference line* show that if the PNIs are not selected in the fist time, they will be selected

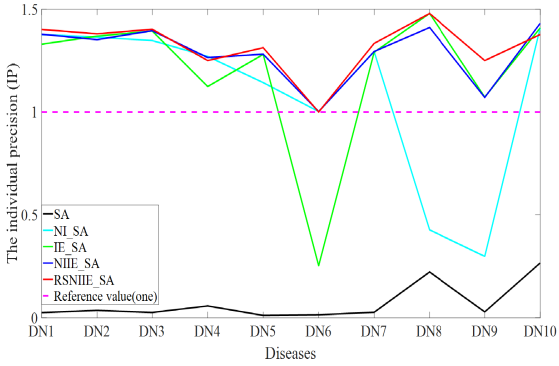


FIGURE 9. The individual precision (*IP*) of five methods.

in the second time. So, positive nutritional ingredients will be selected successfully in either the first or second time. The higher the value is, the greater the chances are for the PNIs to be selected and vice versa. As Fig.9 shows, the effect of using data mining methods is better than using the statistical method called SA. The individual precision based on SA is very low for all diseases. According to Fig.3, it clearly shows that PNIs are not selected in the first several times. The individual precision of other methods are more than one, especially SANIIE_SA, has the highest value for all diseases. For DN6, except for IE_SA, the values of all other methods are around the *reference line*, indicating that the possibility of selecting PNIs successfully in the first time may be very large for other methods. The value of IE_SA is relatively low because DN6 has only two positive nutritional ingredients out of 26, at the same time, PNIs are not selected in the first time.

The methods where *IP* equals to 1.5 indicate that the second positive nutritional ingredient is also selected correctly in the second time, which can be seen from Equation (12). Even though the results show that the *IP* of NIEI_SA is very similar to RSNIEI_SA, the RSNIEI_SA is much better than NIEI_SA from an actual application view. In the case of the same *IP*, the NIEI_SA gives 26 nutritional ingredients, but RSNIEI_SA gives only a few nutritional ingredients, thus, RSNIEI_SA reduces the possibility of wrong recommended food with PNIs. From the view of individual diseases, RSNIEI_SA is the best, because it identifies positive nutritional ingredients correctly in first two times.

As Fig.10 shows, the parts above the *reference line* indicate that the first position are PNIs for all testing samples. As can be also seen from Fig.10, when we plot the graph using *Matlab*, we have only used three decimal points. From Fig.10, SA selects successfully zero positive nutritional ingredients for all diseases. For the NI_SA, the value of *OP* is 1.081, which is almost one, this shows that the PNIs are all selected successfully for all diseases in the first time. In Fig.9, we do not see big performance difference between NIEI_SA and RSNIEI_SA. However, it is quite obvious in Fig.10 that RSNIEI_SA is the best one, because the *OP* value of RSNIEI_SA is 0.5 higher than NIEI_SA. Over all, the value of *OP* using data mining methods is greater than the *OP* value of the statistical SA method, especially RSNIEI_SA,

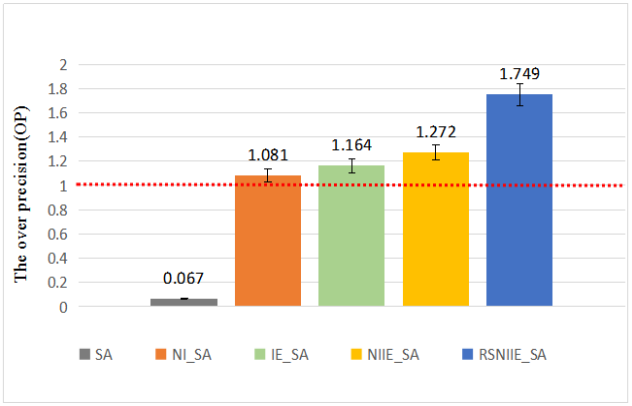


FIGURE 10. The over precision (*OP*) of five methods.

it achieved 1.682 improvements, which shows that it can be applied in the disease analysis based on the mining of nutritional ingredients in food.

VI. CONCLUSION

The main work of this paper can be divided into two parts: firstly, we obtained and sorted out more than seven thousand Chinese diseases and corresponding recommended and taboo food from medical and official websites; secondly, we discussed the relationship between nutritional ingredients and diseases, which mainly aims to find out which ingredients play a positive role in the rehabilitation of diseases. To the best of our knowledge, this is the first study in China, which mines the relationship between nutritional ingredients in food and diseases by using data mining technology. Experimental results showed that although we could not completely find all the positive nutritional ingredients for diseases by data mining methods, the first two or three ones were selected accurately. In addition, if our perspective can be combined with taboo food, the results would be likely to be better and in line with reality, which will be our future work direction.

There are two main benefits of this work: 1) You can access to our website[2] to get diseases, recommended food, taboo food and corresponding nutrition information involved in this paper; 2) We can assist doctors and disease researchers to find out positive nutritional ingredients that are conducive to the rehabilitation of the diseases as accurately as possible. At present, some data is not available, because they are still in the medical verification. Besides, our knowledge base is still gradually improving, if researchers find out something incorrect in our work, we hope to contact us and make our research improved.

REFERENCES

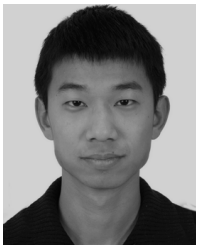
[1] *Global Nutrition Report*, CNS, Beijing, China, 2016.
[2] *Global Status Report on Noncommunicable Diseases*, WHO, Geneva, Switzerland, 2014.
[3] S. Balsari *et al.*, "A retrospective analysis of hypertension screening at a mass gathering in India: Implications for non-communicable disease control strategies," *J. Hum. Hypertension*, vol. 31, no. 11, pp. 750–753, 2017.
[4] *Chinese ResidentâĉĂŹs Chronic Disease and Nutrition (2015)*, DNHFPC of PRC, Beijing, China, 2015.
[5] S. Tellier *et al.*, "Basic concepts and current challenges of public health in humanitarian action," in *International Humanitarian Action: NOHA Textbook*. Cham, Switzerland: Springer, 2017, pp. 229–317, doi: 10.1007/978-3-319-14454-2_13.
[6] F. Ara, F. Saleh, S. J. Mumu, F. Afnan, and L. Ali, "Awareness among Bangladeshi type 2 diabetic subjects regarding diabetes and risk factors of non-communicable diseases," *Diabetologia*, vol. 54, p. S379, Sep. 2011, doi: 10.1007/s00125-011-2276-4.
[7] *Report of Market Prospective and Investment Strategy Planning on China Intelligent Medical Construction Industry (2017-2022)*, QIANZHAN, Forward, 2017.
[8] W. H. Ling, "Progress of nutritional prevention and control on noncommunicable chronic diseases in China," *China J. Dis. Control Prev.*, vol. 21, no. 3, pp. 215–218, 2017, doi: 10.16462/j.cnki.zhjbkz.2017.03.001.
[9] M. B. Margaret, B.-K. Barbara, and D. Colette, "Developing health promotion workforce capacity for addressing non-communicable diseases globally," in *Global Handbook on Noncommunicable Diseases and Health Promotion*. New York, NY, USA: Springer, 2013, pp. 417–439, doi: 10.1007/978-1-4614-7594-1_28.
[10] S. M. Williams and J. H. Moore, "Lumping versus splitting: The need for biological data mining in precision medicine," *BioData Mining*, vol. 8, no. 16, pp. 1–3, 2015.
[11] G. M. Oppenheimer, "Framingham heart study: The first 20 years," *Prog. Cardiovascular Diseases*, vol. 53, no. 1, pp. 55–61, 2010.
[12] W. Y. Jiao, Y. Xue, T. C. He, Y. M. Zhang, and P. Y. Wang, "Association between South Korean dietary pattern and health," *Food Nutrition China*, vol. 23, no. 5, pp. 81–84, 2017.
[13] K. W. Lee and M. S. Cho, "The traditional korean dietary pattern is associated with decreased risk of metabolic syndrome: Findings from the Korean National Health and Nutrition Examination Survey, 1998–2009," *J. Medicinal Food*, vol. 17, no. 1, pp. 43–56, 2014.
[14] H. Du *et al.*, "Fresh fruit consumption and major cardiovascular disease in China," *New England J. Med.*, vol. 374, no. 14, pp. 1332–1343, 2016.
[15] H. Li *et al.*, "A genome-wide association study identifies *GRK*5 and *RASGRP*1 as type 2 diabetes loci in Chinese Hans," *Diabetes*, vol. 62, no. 1, pp. 291–298, 2013.
[16] G. Agapito *et al.*, "DIETOS: A recommender system for adaptive diet monitoring and personalized food suggestion," in *Proc. IEEE 12th Int. Conf. Wireless Mobile Comput., Netw. Commun.*, vol. 4, Oct. 2016, pp. 1–8.
[17] X. L. Zeng, Y. Fu, and H. P. Qing, "Denosing method based on wavelet transform," *Comput. Appl.*, vol. 25, no. 9, pp. 2140–2142, 2005.
[18] G. S. Zhang and X. D. Zhang, "A SVM-GARCH model for stock price forecasting based on neighborhood mutual information," *Chin. J. Manage. Sci.*, vol. 24, no. 9, pp. 11–20, 2016.
[19] C. S. Diogo, A. A. Greta, E. Marcio, A. Romis, and S. Ricardo, "Using an evolutionary denoising approach to improve the robustness of chaotic synchronization," *IFAC Proc. Volumes*, vol. 45, no. 12, pp. 35–39, 2012.
[20] R. Tang, S. Fong, R. K. Wong, and K. K. L. Wong, "Dynamic group optimization algorithm with embedded chaos," *IEEE Access*, vol. 6, pp. 22728–22743, 2018.
[21] J. H. Kuang, "The mean and noise of protein numbers in stochastic gene expression and their dynamical behaviors," M.S. thesis, GuangZhou Univ., Guangzhou, China, 2012.
[22] M. Pájaro, A. A. Alonso, I. Otero-Muras, and C. Vázquez, "Stochastic modeling and numerical simulation of gene regulatory networks with protein bursting," *J. Theor. Biol.*, vol. 421, pp. 51–70, May 2017.
[23] Z. Lei, S. Wang, and D. Xu, "Protein sub-cellular localization based on noise-intensity-weighted linear discriminant analysis and an improved *K*-nearest-neighbor classifier," in *Proc. 9th Int. Congr. Image Signal Process., Biomed. Eng. Inform.*, Oct. 2016, pp. 1871–1876.
[24] K. Yang and H. W. Jiang, "Research of improved algorithm for multilevel thresholding image segmentation based on fuzzy maximum entropy," *Comput. Eng. Appl.*, vol. 45, no. 32, pp. 174–177, 2009.
[25] S. Deb, Z. Tian, S. Fong, R. Wong, R. Millham, and K. K. L. Wong, "Elephant search algorithm applied to data clustering," *Soft Comput.*, vol. 22, no. 18, pp. 6035–6046, Sep. 2018, doi: 10.1007/s00500-018-3076-2.
[26] E. Avci, "Comparison of wavelet families for texture classification by using wavelet packet entropy adaptive network based fuzzy inference system," *Appl. Soft Comput.*, vol. 8, no. 1, pp. 225–231, 2008.

[2] *http://www.pzcnet.com/*

[27] M. A. Bakhshali, "Segmentation and enhancement of brain MR images using fuzzy clustering based on information theory," *Soft Comput.*, vol. 21, no. 22, pp. 6633–6640, 2017.
[28] I. Zitouni and R. Sarikaya, "Arabic diacritic restoration approach based on maximum entropy models," *Comput. Speech Lang.*, vol. 23, no. 3, pp. 257–276, 2009.
[29] L. Tan and D. Taniar, "Adaptive estimated maximum-entropy distribution model," *Inf. Sci.*, vol. 177, no. 15, pp. 3110–3128, 2007.
[30] W. Zhao, H. Liu, W. Dai, and J. Ma, "An entropy-based clustering ensemble method to support resource allocation in business process management," *Knowl. Inf. Syst.*, vol. 48, no. 2, pp. 305–330, 2016.
[31] H. Chen, "Long-life MAODV protocol based on entropy measurement," *Comput. Eng.*, vol. 33, no. 6, pp. 158–160, 2007.
[32] M. Lenič, P. Povalej, and P. Kokol, "Impact of purity measures on knowledge extraction in decision trees," *Found. Novel Approaches Data Mining*, vol. 9, pp. 229–242, Nov. 2005.
[33] A. John, Z. Yang, R. Riahi, and J. Wang, "A decision support system for the assessment of seaports' security under fuzzy environment," *Model., Comput. Data Handling Methodol. Maritime Transp.*, vol. 131, pp. 145–177, Sep. 2017.
[34] C. J. Yang, H. Ge, and Z.-S. Wang, "Overview of attribute reduction based on rough set," *Appl. Res. Comput.*, vol. 29, no. 1, pp. 16–20, 2012.
[35] W. Liu, Y. Gu, Y. Feng, and J. Wang, "An improved attribute reduction algorithm of decision table," *Pattern Recognit. Artif. Intell.*, vol. 17, no. 1, pp. 119–123, 2004.
[36] H. S. Nguyen, "Approximate Boolean reasoning: Foundations and applications in data mining," in *Transactions on Rough Sets V*, vol. 4100. Springer, 2006, pp. 334–506, doi: 10.1007/11847465.
[37] Y. Dong, B. Xiang, T. Wang, H. Liu, and L. Qu, "Rough set-based SAR analysis: An inductive method," *Expert Syst. Appl.*, vol. 37, no. 7, pp. 5032–5039, 2010.
[38] L. Zhang, Y.-M. Ye, S. Yu, and F.-Y. Ma, "A new multivariate decision tree construction algorithm based on variable precision rough set," in *Proc. Int. Conf. Web-Age Inf. Manage.*, vol. 2762, pp. 238–246, 2003.
[39] X. Jia, L. Shang, B. Zhou, and Y. Yao, "Generalized attribute reduct in rough set theory," *Knowl.-Based Syst.*, vol. 91, pp. 204–218, Jan. 2016.
[40] T. Zheng, "An improved MBARK algorithm based on the importance of attribute and mutual information," *Inf. Secur. Technol.*, vol. 8, pp. 11–14, 2012.
[41] L. G. Machado, J. C. C. B. S. de Mello, and M. C. Roboredo, "Efficiency evaluation of Brazilian electrical distributors using DEA game and cluster analysis," *IEEE Latin Amer. Trans.*, vol. 14, no. 11, pp. 4499–4505, Nov. 2016.
[42] J.-Y. Wang, L.-M. Liu, and A. Luo, "Research on decomposition of large decision table," *Sci. J. Comput. Sci.*, vol. 34, no. 8, pp. 211–214, 2007.
[43] J. S. Chen, "Nutrition, physical activity and cancer prevention: A global perspective," Amer. Inst. Cancer Res., Washington, DC, USA, Tech. Rep., 2007.
[44] K. H. Han, "In-band optical signal-to-noise ratio monitoring for suppression of effects of both polarization mode dispersion and the polarization extinction ratio of a polarization beam splitter based on enhanced tracking of the principal states of polarization," *J. Korean Phys. Soc.*, vol. 62, no. 6, pp. 897–901, 2013.
[45] F. Xie, J. G. Gu, and Z. W. Lin, "Assessment of aquatic ecosystem health based on principal component analysis with entropy weight: A case study of Wanning Reservoir (Hainan Island, China)," *Chin. J. Appl. Ecol.*, vol. 25, no. 6, pp. 1773–1779, 2014.
[46] S.-F. Ding, J. Zhang, X.-K. Zhang, and Y.-X. An, "Survey on multi class twin support vector machines," *J. Softw.*, vol. 29, no. 1, pp. 89–108, 2018.
[47] Z. Pawlak, *Rough Sets: Theoretical Aspects of Reasoning about Data*. Dordrecht, The Netherlands: Academic, 1991.
[48] S. Wang, "Rough set theory and knowledge acquisition," M.S. thesis, Xi'an Jiaotong Univ., Xi'an, China, 2001.
[49] W. J. Liu, "An attribute reduct algorithm of continuous domain decision table," *J. East China Univ. Sci. Technol. (Natural Sci. Ed.)*, vol. 33, pp. 13–16, 2007.
[50] E. Matthes and G. Z. Yuan, *Python Crash Course: A Hands-On, Project-Based Introduction to Programming*. San Francisco, CA, USA: Posts & Telecom Press, 2016.
[51] C. Shi, *Selenium 2 Test Automation Practices-Based on the Python Language*. Beijing, China: Publishing House Electron. Ind., 2016.
[52] Y. X. Yang, G. Y. Wang, and X. C. Pan, *China Food Composition*. Beijing, China: Peking Univ. Med. Press, 2002.
[53] Y. X. Yang, G. Y. Wang, and X. C. Pan, *China Food Composition*, 2nd ed. Beijing, China: Peking Univ. Med. Press, 2004.
[54] F. P. Miller, A. F. Vandome, and J. McBrewster, *Binary-Coded Decimal*. VDM Publishing House, 1988.
[55] E. Xiong, C. Zheng, X. Wu, and W. Wang, "Protein subcellular location: The gap between prediction and experimentation," *Plant Mol. Biol. Reporter*, vol. 34, no. 1, pp. 52–61, 2016.

ZHENFENG LEI received the B.S. degree in computer science from Zhengzhou University in 2015 and the M.A.Eng. degree (Hons.) in computer science from Yunnan University in 2017. He is currently pursuing the Ph.D. degree with the School of Information Science and Engineering, Xiamen University, Xiamen, China. He had special insights in the field of protein sub-cellular localization. His current research interests include data mining and deep learning techniques, knowledge graph, and recommended system. He won the First Prize of the China Graduate Contest on Application, Design and Innovation of Mobile-Terminal, which held at Xiamen University in 2018.

SHUANGYUAN YANG received the Ph.D. degree from the Computer School, Huazhong University, Wuhan, China, in 2004. He is currently an Associate Professor with the Software School, Xiamen University, Xiamen, Fujian, China. In the past five years, he has hosted or participated in many projects, including two National Natural Science Foundation Projects, one National 863 Project, one Major Science and Technology Project in Fujian Province, five Key Science and Technology Projects in Fujian Province and Xiamen City, and over 10 Enterprise Cooperation Projects. Among them, as the person in charge of the project, nine projects were undertaken, with a total amount of 7.93 million RMB. He has published over 30 journal or conference papers. His research topics include Internet of Things, SAAS service, image processing, and machine learning. He received the honorary title of Introducing High-Level Software Talents in Fujian Province in 2010 and the third class of Outstanding Talents in Jinjiang City in 2011.



HAN LIU (S'15–M'16) received the B.Sc. degree in computing from the University of Portsmouth in 2011, the M.Sc. degree in software engineering from The University of Southampton in 2012, and the Ph.D. degree in machine learning from the University of Portsmouth in 2015. He was a Research Associate in computational intelligence with the School of Computing, University of Portsmouth. He is currently a Research Associate in data science with the School of Computer Science and Informatics, Cardiff University. He has authored over 40 papers in the areas such as data mining, machine learning, and granular computing. He published a research monograph with Springer in the third year of his Ph.D., and his another monograph was published in 2018. One of his papers was identified as a key scientific article contributing to scientific and engineering research excellence by the selection team at Advances in Engineering, and the selection rate is less than 0.1% as indicated. He also has a paper selected as the Finalist of the Lotfi Zadeh Best Paper Award in the 16th International Conference on Machine Learning and Cybernetics (ICMLC 2017) and has another paper nominated for Lotfi Zadeh Best Paper Award in the 15th International Conference on Machine Learning and Cybernetics (ICMLC 2016). His research interests include data mining, machine learning, rule-based systems, granular computing, intelligent systems, fuzzy systems, big data, computational intelligence and applications in cyber security, cyber crime, cyber bullying, cyber hate, and pattern recognition.

SABA ASLAM (M'14) received the B.S. degree (Hons.) in software engineering from GCUF, Pakistan, and the master's degree in computer technology from Xiamen University, China. She is currently a Researcher, who has strong expertise in data mining, artificial intelligence, and machine learning. She has experience in academia as a Lecturer and also in the development field. She is currently focusing on Stock Exchange Market Forecasting using deep learning. She has excellent interpersonal skills, keen problem solving, and analytical skills essential for an enterprising engineering professional. She is a Bronze Medalist. She is the Chair of the IEEE-WIE GCUF, Pakistan.

JINYU LIU received the B.S. degree in computer science and technology from Xiamen University in 2018. She is currently pursuing the master's degree with the School of Engineering, The Hong Kong University of Science and Technology. Her interests include well application of data analysis method in specific industry. Her research interests include financial mining and image processing. She won the First Prize of the China Undergraduate Mathematical Contest in Modeling in Fujian Province.

HALEFOM TEKLE received the B.Sc.Eng. degree in information technology from the Mekelle Institute of Technology and the M.Tech. degree in computer and information technology from Defence University College of Engineering, Ethiopia, in 2009 and 2014, respectively. He is currently pursuing the Ph.D. degree with the School of Information Science and Engineering, Xiamen University, Xiamen, China. His current research interest is in optical character recognition and machine translation using deep learning techniques.

EMMANUEL BUGINGO received the B.Sc. degree (Hons.) in information and communication technology from the University of Rwanda (Formal Umutara Plytechnic), Rwanda, in 2012, and the master's degree in computer science from Xiamen University, China, in 2016, where he is currently pursuing the Ph.D. degree with the Department of Computer Science. His research interests include computer vision for image mining using deep learning, cloud computing, big data-processing, and workflow scheduling.

DEFU ZHANG (M'16) received the bachelor's and master's degrees in computational mathematics from Xiangtan University in 1996 and 1999, respectively, and the Ph.D. degree in computer software and its theory from the School of Computer Science, Huazhong University of Science and Technology. He was a Senior Researcher with Shanghai Jinxin Financial Engineering Academe from 2002 to 2003. He was a Post-Doctoral Researcher with the Longtop for Financial Data Mining Group from 2006 to 2008. From 2008 to 2016, he visited Hong Kong City University, University of Wisconsin–Madison, and Macau University. Besides, he developed an Internet + big data platform (http://www.pzcnet.com). He is currently a Professor with the Department of Computer Science, Xiamen University. He supervised the ACM/ICPC Team, Xiamen University. He has authored over 40 journal articles. His research interests include computational intelligence, data mining, big data, cloud computing, online decision optimization, and food security. He was a recipient of three gold medals , and eight silver medals from 2004 to 2009, and he took part in the World Final Contest in 2007.