Acta Technica *62* (2017), No. 5A, 957–964

Attribute reduction of macro genomic gene signal based on improved rough set reduction algorithm

Jian Xue[2,3], Wang Zhongli[3,4]

Abstract. When K-mer frequency is used as a digital feature of DNA fragments, it is time-consuming and computationally expensive to classify it. If the K-mer frequency is effectively reduced before classifying, the remaining attributes can be kept of all the information on behalf of the original digital features, it will greatly improve the efficiency of classification. In this paper, we propose a method of attribute reduction based on rough set theory and Selpso algorithm. The experimental results show that this method can effectively remove redundant attributes and improve classification efficiency.

Key words. Metagenomics, metagenomics, rough set and selpso attribute reduction, feature classification.

1. Introduction

There are about 99% microorganisms can not be detect by molecular biology techniques. The metagenomics technology (Metagenomic) refers to a direct sequencing of microbial DNA in environmental samples all the information method[1,2]. Currently the technology has been widely used in soil, deep sea, extreme ecological environmen[3–5], intestinal disease in different patients[6–8]and so on. In 1998, Handelsman et al. [9]created the meta genomic method to directly extract the DNA of all the microorganisms in the soil sample. Among them, Contains a large number of previously unknown and cloned genes, direct sequencing of a small ecological environment samples of all micro-organisms in the DNA method for meta genomics

[1] Acknowledgement - This work is supported by Key technology public relations project of science and Technology Department, Jilin Province. Project number: 20170204021GX, Name: Intelligent running fitness platform based on the real street view.Key scientific and technological research projects.

[2] Workshop 1 - College of Communication Engineering, Jilin University, Chang Chun, China

[3] Workshop 2 - College of Electrical & Information Engineering, Beihua University, Jilin, China

[4] Corresponding author: Zhongli Wang

technology.[10]At present, there are two categories bioinformaticians platform used for authenticating the metagenome DNA fragments: one is based on the authentication of DNA sequences characters, such as BLAST , CARMA, MEGAN, TreePhyler , MetaDomain , etc.This platform can't authenticate effectively when the database has no source genome of DNA sequences. The other kind is based on the DNA digital characteristics which can overcome the narrow scope shortcoming mentioned above. Such as: PhyloPythia, Phylopythias, TACOA, NBC, PhymmBL .

Figure 1 shows the general meta genomics technology flow. With k-mer frequency as a species tag, in order to achieve better recognition accuracy, we need to extract a vector of nearly a thousand dimensions can be achieved. In the whole process of gene signal classification and identification, it is effective to reduce the vector of high dimension and improve the speed of identification under the premise of guaranteeing the accuracy of classification has become a research value. In this paper, a method of attribute reduction based on the combination of natural selection particle swarm optimization and rough set theory is studied. Experiments are carried out on the whole genome sequences of 20 bacteria. The results show that the method can remove the redundant attributes, The attribute still can be very good representative of all the original bacterial properties of all information

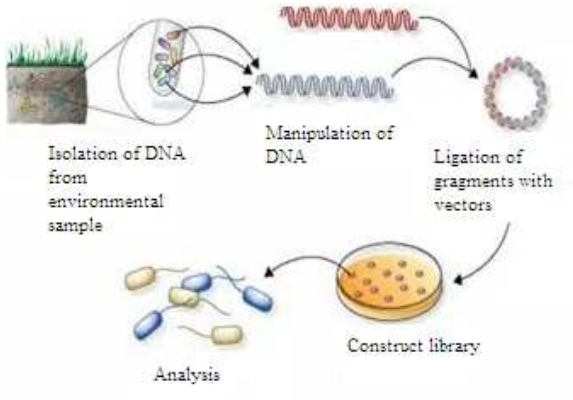

Fig. 1. The technical process of meta genomics

2. rough set theory and selpso algorithm attribute reduction process

Before reducing the K-mer frequency, the K-mer frequencies to be reduced are first grouped into the decision table according to the knowledge system. The form of the decision table is shown in Table 1. The first column is the sample object label, the last column is the category of each sample, and the middle is the attribute value of the sample. Here, all values in the table are required to be discrete. The establishment of attribute decision table is the primary task of rough set attribute reduction. The form of a complete decision table is as follows.

Decision table is an information table knowledge expression system,$S =< U, R, V, f >$,$R = C \bigcup D$is the attribute set,Subset $C$and $D$are called conditional and result attribute sets, respectively.Equivalence relation between condition attribute C and result attribute D Equivalent class of$IND(C)$and $IND(D)$ is called condition class and decision class respectively.The following steps are required to obtain the meta genomic attributes in Table 1.

Table 1. Decision table form

Number	Attribute1	Attribute 2	...	Attribute m	classify
1	c_{11}	c_{12}	...	c_{1i}	d_1
2	c_{21}	c_{22}	...	c_{2i}	d_2
3	c_{31}	c_{32}	...	c_{3i}	d_3
⋮	⋮	⋮	⋮	⋮	⋮
n	c_{n1}	c_{n2}	...	c_{ni}	d_n

A. Genomic source code acquisition

The experimental data in this paper are derived from the American Center for Bioinformatics

(NCBI, http://www.ncbi.nlm.nih.gov/).

The example data information obtained on the website is shown in Figure 2.

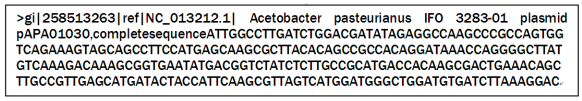

Fig. 2. DNA of Acetobacter_pasteurianus_IFO_3283_01_uid59279

B. Interception Isometric Sequences and K-Mer Frequency Acquisition

Fragments of different species are different in length ,in order to avoid the effect of DNA fragment length on the feature length, the original sequence should be divided into equal lengths before calculating the K-mer frequency. Herein, the DNA fragment is divided into segments each 1000 bp (i.e., 1000 nucleotides), and 50 segments are randomly selected as the experimental samples.

Purines in adenine (Adenine, A) and guanine (G) are pyrimidine cytosine (C) and thymine (Thymine, T). We can think of DNA sequence is A, T, C, G four "letters" arranged by a particular law of a long string of letters sequence, which is the object we want to study. It has shown that the frequency of a short sequence fragment is stable throughout the genome and that this distribution is conserved and representative. Zhou found that the genome of all organisms, the length of 500 bp to 10000 bp DNA fragments have a stable 4-mer frequency distribution.Figure 3 shows the distribution of the three bacteria belonging to the same species and three genera. It can be seen from the figure that the distribution of the 4-mer from the same genus is very similar, but from the different 3 the frequency of bacterial 4-Mer

of the genera is very different.

A total of 4k different "words" can be obtained ,in the "text" detect of each "word" and the number of occurrences of its normalized, we can get a dimension of 4^k eigenvector for identification, later call this DNA number character “k-mer frequency”.According to the definition of k-mer frequency, we assign A = 0, T = 1, C = 2, G = 3 to the four nucleotides constituting the DNA sequence, the gene signal is digitized to 0, 1, 2,3, Four figures according to a certain biological relationship to form data The k-mer frequency can be expressed as Equation 1:

$$\boldsymbol{F}(k) = \frac{c_{ki}}{\max(c_{ki})} \tag{1}$$

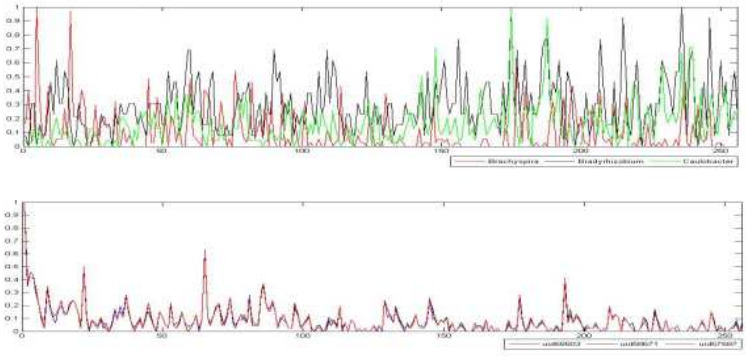

Fig. 3. 4-mer bacteria distribution

Where c_{ki} is the number of occurrences of different combinations of permutations of "string" of length k, $\max(c_{ki})$ is the largest among them,$i = 0, 1, 2, \ldots, 4^k - 1$,The result is an eigenvector of length k,After this treatment, the abscissa of Figure. 1 successively represents AAAA, AAAT, AAAC, ...,GGGC, and GGGG in 256 features,The ordinate is the frequency at which each feature appears, and its population is theK-mer frequency eigenvector of the species.

C. Discretization of K-mer Frequencies

As the rough set can not deal with continuous vector, so after the K-mer frequency normalization process, we have to discretize it, Otsu algorithm can calculate the various gray levels of inter-class variance to find the optimal cut off Points, thereby discretizing the gray-scale data. In this paper, we use the Otsu algorithm for attribute discretization.

The maximum value of the attribute to be reduced is $x_{\max} = \max(x_1, x_2, \ldots, x_N)^{\mathrm{T}}$, ,The minimum value is $x_{\min} = \min(x_1, x_2, \ldots, x_N)^{\mathrm{T}}$,The range of the attributes to be reduced is $[x_{\min}, x_{\max}]$,The interval was divided into L portions.The probability and the mean value of the fall on each interval are shown in Eqs. (2) and (3):

$$P_i = \frac{N_i}{N} \quad i = 1, 2, \ldots, L \tag{2}$$

$$\mu_i = \frac{\sum_0^{N_i} u_i}{N_i} \quad i = 1, 2, \ldots, L \tag{3}$$

Where $P_i$is the probability that the attribute value falls within the$i-$th interval, $N_i$is the number of samples falling in the$i-$th interval, μ_i is the mean value of the attribute that falls in the i-th interval. The formula for the variance between classes is shown in Eq. (4):

$$g = \sum_{i=1}^{L} P_i(\mu_i - \mu)^2 \quad i = 1, 2, \ldots, L \tag{4}$$

Where μ is the total mean of attributes,$\mu = \frac{\sum_{j=1}^{N} x_j}{N} \quad j = 1, 2, \ldots, N$, When the interclass variance to obtain the maximum value, it corresponds to the property value, that is the breakpoint.

D. Feature reduction

Combining the natural selection mechanism with the PSO algorithm, a PSO algorithm based on selection is obtained, The basic idea is to sort the whole particle swarm by fitness value in each iteration, replacing the worst half position and velocity with the velocity and position of the best half of the population, while preserving the memory of each individual Historical optimal value.

The attribute set is an independent variable, Set the corresponding amount of information corresponding to the size of the function of the value to find the optimal value of this function, so you can Selpso particle swarm optimization algorithm to solve .

Assuming that A is a property set, A is the value of an attribute in the collection. If there are two samples X_i, X_j, the attribute values are the same under$\forall a \in A, A \subset R, X_i \in U, X_j \in U$ conditions, That is$f_a(X_i) = f_a(X_j)$ is established, Then the object X_i and X_j is said to be equivalent to the property A relationship. The two samples with the same attribute value can be called equivalence relations. The equivalence relation can be expressed by Eq.(5).

$$IND(A) = \{(X_i, X_j) | (X_i, X_j) \in U \times U, \forall a \in A, f_a(X_i) = f_a(X_j)\} \tag{5}$$

In the domain U, The equivalent set $[X]_A$ is the set of elements in the attribute set A that have the same equivalence relation as $IND(A)$, Can be expressed by Eq.(6):

$$[X]_A = \{X_j | (X, X_j) \in IND(A)\} \tag{6}$$

From the above formula, we can see that the object with the same attribute value constitutes the equivalent set, The equivalence partition of all equivalence sets of attribute A can be expressed by Eq. (7)

$$A = \{E_i | E_i = [X]_A, i = 1, 2, \cdots\} \tag{7}$$

In this paper, the fitness function is chosen in two parts,On the one hand, it is

necessary to make the remaining condition attributes after reduction less than 256 before reduction, on the other hand, to ensure that the attribute dependency of remaining attributes after reduction is large.The objective function is composed of two parts, defined as

$$F(r) = \gamma_B(D) + {}^{m - m_r}/_{m} \tag{8}$$

$\gamma_B(D)$represents the attribute dependency,$m_r$denotes the number of attributes is 1, and $m$denotes the number of attributes.

Given a decision system$S$, $\forall B \subseteq C$,The dependency of decision attribute B on condition attribute subset D is

$$\gamma_B(D) = \frac{|Pos_B(D)|}{|U|} \tag{9}$$

$|POS\,(C,D)|$and$|U|$ are the number of objects in the positive field$POS\,(C,D)$ and the number of objects in the whole object.

3. experimemt Results

A 4-mer (i.e., k = 4) frequency was extracted from each DNA fragment as the initial gene signal. After the data preprocessing, we apply the rough set and the Selpso PSO algorithm to the 4-mer frequency reduction. The reduced gene

signals were compared with Weka classification platform to identify the microbial DNA fragments before and after reduction. As can be seen from the compares result shown in Table 2, the original data 256-dimensional 4-mer feature vector reduction after the remaining attributes of 68 and 61 dimension, reducing the 188 and 195-dimesion, the ratio of 73.43% and 76.17%,And after the reduction the feature attributes in the classification accuracy rate has not declined but a small increase. It is shown that the redundant attribute can be effectively removed by the reduction method, and the feature of the feature can be extracted effectively, and the efficiency of classification can be improved.

Table 2. Comparision of the indicators before and after reduction

	The Number of Feature Attributes before Reduction	The Number of Feature Attributes after Reduction	Pre - reduction Classification Accuracy	Accuracy of classification after reduction	Test number	The exact number before reduction	The exact number after reduction
10 randomly selected genomes	256	68	92.78%	93.89%	2	167	169
20 randomly selected genomes	256	61	90.93%	92.22%	12	983	996

4. Conclusions

In this paper, rough set theory and Selpso particle swarm optimization algorithm were used to reduce the genomic DNA fragments, The experimental results showed that the method reduced the feature dimension while keeping or small amplitude to improve classification accuracy. the experiments in this paper are carried out at the level of “genus” macro genomics, and further experiments should be carried out at the level of "species" and obtain effective classification after reduction Precision.

References

[1] J. Handelsman, M. R. Rondon, S. F. Brady: *Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products.* Chem Bio *10* (1998), No. 5, 245.

[2] P. J. Turnbaugh, J. I. Gordon: *An invitation to the marriage of metagenomics and metabolomics.* Cel *5* (2008), No. 134, 708.

[3] S. R. Gill, M. Pop, R. T. Deboy, P. B. Eckburg, P. J. Turnbaugh: *Metagenomic Analysis of the Human Distal Gut Microbiome.* Science *5778* (2006), No. 312, 1355.

[4] F. Warnecke, P. Luginbühl, N. Ivanova: *Metagenomic and functional analysis of hind gut microbiota of a wood-feeding higher termite.* Nature *7169* (2007), No. 450, 560.

[5] M. L. Sogin, H. G. Morrison, J. A. Huber: *Microbial diversity in the deep sea and the underexplored "rare biosphere".* Proc Natl Acad Sci *32* (2006), No. 103, 12115.

[6] D. B. Rusch, A. L. Halpern, G. Sutton: *The Sorcerer II Global Ocean Sampling Expedition: Northwest Atlantic through Eastern Tropical Pacific.* PLoS Biol *3* (2007), No. 3, 77.

[7] C. L. Hemme, Y. Deng, T. J. Gentry: *Metagenomic insights into evolution of a heavy metal-contaminated groundwater microbial community.* ISME J *5* (2010), No. 4, 660.

[8] A. Manimozhiyan, R. Jeroen: *Enterotypes of the human gut microbiome.* Nature (2011) 174.

[9] J. Handelsman, M. R. Rondon, S. F. Brady: *Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products.* Chemistry & biology *1* (1998), No. 5, 242–249.

[10] P. J. Turnbaugh, J. I. Gordon: *An invitation to the marriage of metagenomics and metabolomic.* Cell *5*, (2008), No. 134, 708–713.

[11] S. F. Altschul, W. Gish, W. Miller, E. W. Myers: *Basic local alignment search tool.* J Mol Biol *3* (1990), No. 215, 403–410.

[12] L. Krause, N. N. Diaz, A. Goesmann: *Phylogenetic classification of short environmental DNA fragments.* Nucleic Acids Res *7* (2008), No. 36, 2230–2239.

Received November 16, 2017