Cellular and Molecular Biology
E-ISSN : 1165-158X / P-ISSN : 0145-5680
www.cellmolbiol.org

Original Research

Design of abnormal data detection system for protein gene library based on data mining technology

Cuixia Liu[1*], Yuwei Wang[2]

[1] Information Department, Shijiazhuang Vocational College of Finance & Economics, Shijiazhuang 050061, China
[2] Global Media Communication, The University of Melbourne, Melbourne 3006, Australia

[*]**Correspondence to:** Louiser75@163.com

Received March 11, 2020; **Accepted** September 31, 2020; **Published** October 31, 2020

Doi: http://dx.doi.org/10.14715/cmb/2020.66.7.16

Abstract: In view of the shortcomings of the current abnormal data detection system of the protein gene library, such as low detection rate and high error detection rate, the abnormal data detection system of the protein gene library based on data mining technology is designed. The protein gene enters the firewall module of the system, and enters the immune module when it does not match the firewall rules; the memory detector in the immune module presents the protein gene, if the memory detector does not match the protein gene, the mature detector presents the protein gene, if the mature detector does not match the protein gene, it is determined as the normal protein gene data package, if it matches, it is considered that The abnormal data of protein gene was processed by the collaborative stimulation module, and the control module controlled by C8051F060 chip to detect the abnormal data of protein gene library. The immune module generates new protein gene sequences through an immature detector, simulates the immune mechanism of protein gene through a mature detector module, and simulates the secondary response in the abnormal data detection system of protein gene library through memory detector. The system introduces data mining technology into the detection and uses a two-level dynamic optimization algorithm to calculate the ASG similarity value of protein gene secondary structure arrangement. According to this value, the abnormal data detection of the protein gene library is realized by randomly generating protein genes, negative selection, clone selection and copying memory cells through gene expression. The experimental results show that the system can quickly detect abnormal data of the protein gene library, ensure the detection efficiency, and the detection accuracy reaches 97.1%. The system can reduce the error rate of normal protein gene detection as an abnormal protein gene.

Key words: Data Mining Technology, Protein Gene Library, Abnormal Data, Detector, Accuracy, Detection System.

Introduction

GenBank is an authoritative sequence database established and maintained by NCBI, the U.S. biotechnology information center. It is one of the largest databases of nucleotide and protein gene sequences in the world (1). It collects all the published DNA and protein gene sequences, as well as the relevant biological information and references. The gene data in the gene pool come from more than 47000 different species, including more than 8 billion bases. Now more than 600 new species are added to the database every month, and the data in the gene database doubles every 14 months, and there is a trend of acceleration (2). As of August 1999, GenBank has collected about 4610000 sequences with a length of 3400000000 bases, 57% of which are human (Homo sapiens, 49% of which are human ESTs), in addition to DNA sequences of nematodes (C.elegans, 9%), yeast (s.cereversiae), Mus musculus and other organisms. In GenBank, the branch database dbEST and DBSTs are developing the most rapidly. Each data in GenBank contains the accurate description of the sequence, the scientific name and tree classification of the sequence source organism, as well as the characteristic data column. It provides the protein-coding area of the sequence and the sites with special biological significance, such as transcription units, sites or modifications and repeats. It also provides specific sequences of the protein gene.

GenBank has cooperated with EMBL nucleic acid sequence internationally since its establishment (3). In 1987, the National Institute of the genetics of Japan established the DNA data of Japan (DDBJ) and joined the international cooperation between GenBank and EMBL. Now, the three databases collect the nucleic acid sequence information of their respective regions, forming the international nuclear sequence database collaboration, and realize data sharing, and exchange the new sequence records established by their respective databases every day (4). GenBank has established an integrated database (ID) with other nucleotide sequence databases, such as EMBL, DDBJ, GSDB, LANL, etc., and well-known protein gene databases, such as swiss-prot, PIR, PRF and PDB.

The abnormal data detection of the protein gene library is proposed based on biological immune theory. Forrest et al. first applied the abnormal data detection of the protein gene library to the field of computer big data mining. The artificial immune system is studied and concerned by many scholars because of its characteristics of distribution, self-organization, robustness and adaptability. At present, the abnormal data detection technology of protein gene library based on artificial

immune theory mainly focuses on the research of negative selection algorithm, clonal selection algorithm and immune genetic algorithm, among which the negative selection algorithm proposed by Forrest plays an extremely important role (5), which lays a theoretical foundation for the research of abnormal data detection of protein gene library in the field of big data mining, and many subsequent calculation methods are put forward based on negative selection algorithm. Because there are many kinds of protein gene pool, and the number of protein genes is huge, it may not be found in time when the protein gene is abnormal. The abnormal protein gene will lead to the abnormal growth of animals and plants, even death, so it is of great significance to detect the abnormal data of the protein gene pool. Although the protein gene detection system designed in literature (6) has detected the abnormal data of protein gene, its detection rate is not high; although the detection rate of the system in literature (7) is good, it also classifies the normal gene as the abnormal protein gene, so there are disadvantages in terms of the error detection rate; although the abnormal number of protein gene is designed in literature (8), the system in literature (7) has a good detection rate according to the monitoring system. However, big data mining technology has not been cleverly integrated into the system. In recent years, a large number of studies have shown that big data mining technology is the best way to detect abnormal data in the gene library, and also the most accurate detection method. This literature does not use big data mining technology, so the accuracy is not high. In order to avoid the problems in the above literature, this paper designs the abnormal data detection system of the protein gene library based on data mining technology. The system has a high detection rate and low false detection rate when detecting the abnormal data of the protein gene library, and it can effectively detect the abnormal data in the protein gene library.

Materials and Methods

Structure design of protein gene anomaly detection system based on data mining technology

The abnormal data detection system of the protein gene library based on data mining technology mainly includes a collaborative stimulation module, immune module, random generation detector module, gene library module, control module and a firewall module. The system structure is shown in Figure 1.

The working principle of the system was as follows: in terms of the detector, randomly generate detector module to initialize immature detector; an immature detector was negated by self-set, an immature detector that did not match with self-set was transformed into mature detector through tolerance, while immature detector that matched with self-set dies, delete it (9); mature detector that reached activation threshold in the life cycle. The device was converted into a memory detector. In the immune module, calling the detector was more effective. The system updated dynamically the unactivated mature detector and the memory detector which had not been reused at a certain time transferred them to the gene library module, and generated a new immature detector through the multi-point gene mutation recombination algorithm and the random generation detector module, thereby increasing the utilization rate of the detector gene. To ensure the diversity of the immature detector, improved its passing tolerance probability and improve the detection performance (10). In the aspect of anomaly detection, the packets (protein genes) entering the anomaly data detection system first entered the firewall module. If they match the firewall rules in the firewall module, deleted the packets. Otherwise, enter the immune module of the network system; the memory detector presented the protein genes, if they match, deleted them. Otherwise, the mature detector will submit; if the maturity detector fails to match the protein gene after it was submitted (11), it will be judged as a normal protein gene data package and allowed to access the network system. Otherwise, it will be asked for collaborative processing by the collaborative stimulation module. The collaborative stimulation module will analyze the protein gene data package. If it was considered as an abnormal data package, it will be deleted and the corresponding maturity detector matching degree will be determined. When it reached the activation threshold, it will be transformed into a memory detector. If it was considered as a normal protein gene data package, then the matched mature detector antibody will be deleted. The control module controls the immune module, the random generation detector module and the cooperative stimulation module at the same time.

Immune module

The immune module is the main part of the abnormal data detection system of the protein gene library. Almost all protein gene data are detected in the immune module, including the memory detector submodule R, mature detector submodule M, immature detector submodule I.

(A) Immature detector module: in the generation process of an immature detector, firstly, the random generation detector module initializes the immature detector according to the self-set. After a period of time, the gene library module uses the multi-point gene mutation recombination algorithm to generate immature detector (immature detector generates TCP protocol detector, UDP protocol detector and ICMP according to the rules Protocol detector), which can effectively supplement the immature detector to ensure the high efficiency and diversity of the detector in the immune module. The algorithm of multi-

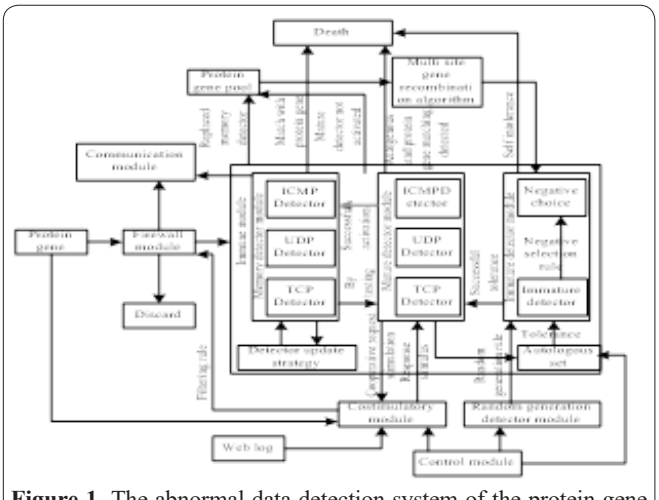

Figure 1. The abnormal data detection system of the protein gene pool.

site gene mutation and recombination is as follows: (a) select the protein gene sequence of different sites randomly from the gene sequence of the autogenous protein gene sample, then there is $A_{g_{i,j}} = resoleve\left(n, m, A_{g_i}\right)$, where $A_{g_{i,j}}$ represents the j protein gene sequence of the i protein gene segment, resoleve function represents the extraction of the protein gene sequence, n represents the position of the extracted protein gene sequence in the protein gene A_{g_i}, indicates the length of the extracted protein gene sequence; (b) the extracted protein gene sequence is combined according to the multi-site gene recombination rule (12); at the same time, the extracted protein gene sequence is mutated (Gauss mutation strategy is adopted here, the Gauss mutation can produce the mutation near the original detector, which has strong local search ability and can retain the original Population information of the initial detector) to generate a new protein gene sequence. It not only keeps the original protein gene type but also makes the new protein gene have diversity (13), which ensures the efficiency of abnormal data detection of the protein gene.

(B) Mature detector module: mature detector module simulates the immune mechanism of the protein gene, which is an important part of detecting the protein gene, and its integrity directly affects the accuracy of detecting abnormal data of the protein gene.

(C) Memory detector module: the memory detector simulates the secondary response in the abnormal data detection system of protein gene library, which is mainly evolved from the mature detector that reaches the activation threshold. When the system is attacked by the same or similar abnormal data of the protein gene again, it can respond quickly, improve the detection efficiency and maintain the stability of the system (14).

Control module

The hardware circuit adopts a C8051F060 control chip, which has powerful functions. Its main features are high-speed, pipelined 8051 compatible CIP-51 core (up to 25mips); two 16 bit, 1 msps ADC with DMA (direct memory access) controller; rich digital I/O pins (59); 4352 (4 K + 256) bytes on-chip RAM; 64KB flash memory programmable in the system; 5 universal 16-bit timers; two UART serial interfaces, etc. The control module is regarded as a control system, and the hardware structure of the control system is shown in Figure 2.

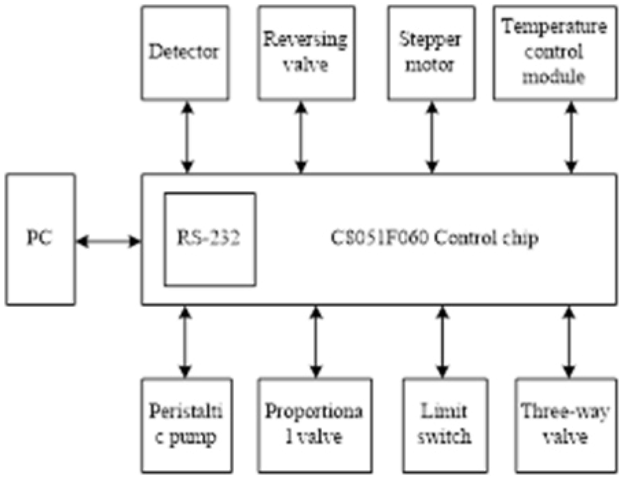

Figure 2. The hardware structure of the control system.

Among them, the analog output of the detector is converted into digital through ADC0 (analog to digital 0 channel) of C8051F060, and the DMA controller integrated with C8051F060 is used to directly read the data into the memory of the single-chip microcomputer. DMA is a data exchange mode that directly accesses data from memory without CPU (15). Its advantage is that it does not need the interference of CPU in data transmission, and it can greatly improve the working efficiency of the CPU. DMA interface works with ADC0 and ADC1 to write ADC output directly to the designated XRAM (expanded RAM) area. The DMA interface is configured by using a special function register. The instruction buffer (16) is accessed through DMA control logic to obtain data from the ADC and control the writing of data to xram. The DMA instruction tells the DMA control logic which ADC to read from but does not start the ADC conversion. DMA control flow is shown in Figure 3.

Feature data extraction

Classification number set $D = \{(a,b,s,o) | a \in A, b \in B, s \in S, o \in O\}$, where A is the classification number set of class layer, B is the classification number set of architecture layer, s is the classification number set of topology layer, and O is the classification number set of the harmony layer.

Mathematical model for ASG calculation

ASG (assignment of protein secondary structure information group) code represents the details of protein gene secondary structure arrangement. Including the name of residue in secondary structure, protein gene chain identifier, PDB residue number, the full name of secondary structure, Phi angle, Psi angle, the contact area of residue solvent and other details (17).

Two string sequences $W_1 = ASSGCASCGCS$, $W_2 = ASCASGSGC$, set $|W_1|$ and $|W_2|$ as their length respectively: $|W_1| = 11$, $|W_2| = 9$, then through the pattern Hunter algorithm, the optimal ratio is as follows:

ASSGCAS— CGC

AS— CASGS GC

The best corresponding set is:

$$n_{W_1W_2} = \{(1,1),(2,2),(5,3),(6,4),(7,5),(10,8),(11,9)\} \quad (1)$$

Two levels of the dynamic optimization algorithm were used to calculate ASG similarity. The process is as follows:

(A) Take the structure data in ASG data as the object for sequence alignment. By using the PatternHunter algorithm, the optimal ratio of the structure strings of two protein genes, A_{CSG_1,CSG_2}, is obtained, and one of the best

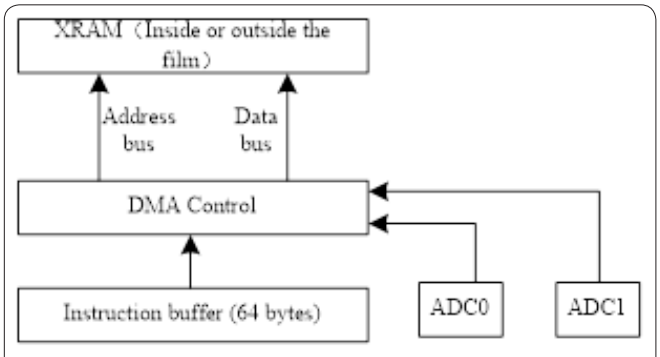

Figure 3. DMA control flow chart.

corresponding sets is obtained

$$n_{W_1W_2} = \{(1,1),(2,3),...,(i,j),...,(n,m)\} \quad (2)$$

(B) Calculate Phi_{CSG_1,CSG_2}, Psi_{CSG_1,CSG_2}, $Area_{CSG_1,CSG_2}$, and PPA_{CSG_1,CSG_2}, respectively:

$$Phi_{CSG_1,CSG_2} = \sqrt{\left(Phi_{e1} - Phi_{f1}\right)^2 + \left(Phi_{e2} - Phi_{f3}\right)^2 + ... + \left(Phi_{ei} - Phi_{fj}\right)^2 + ... + \left(Phi_{ex} - Phi_{fy}\right)^2} \quad (3)$$

$$Psi_{CSG_1,CSG_2} = \sqrt{\left(Psi_{e1} - Psi_{f1}\right)^2 + \left(Psi_{e2} - Psi_{f3}\right)^2 + ... + \left(Psi_{ei} - Psi_{fj}\right)^2 + ... + \left(Psi_{ex} - Psi_{fy}\right)^2} \quad (4)$$

$$Area_{CSG_1,CSG_2} = \sqrt{\begin{array}{l}\left(Area_{e1} - Area_{f1}\right)^2 + \\ \left(Area_{e2} - Area_{f3}\right)^2 + \\ ... + \left(Area_{ei} - Area_{fj}\right)^2 + \\ ... + \left(Area_{ex} - Area_{fy}\right)^2\end{array}} \quad (5)$$

$$PPA_{CSG_1,CSG_2} = k \times Phi_{CSG_1,CSG_2} + n \times Psi_{CSG_1,CSG_2} + t \times Area_{CSG_1,CSG_2} \quad (6)$$

Among them, Phi_{e1} represents the dihedral angle θ formed between the backbone atoms of protein gene e, Psi_{e1} represents another dihedral angle θ formed between the backbone atoms of protein gene e, and $Area_{e1}$ represents the soluble molecular surface area of protein gene e. PPA_{CSG_1,CSG_2} represents the sum of mean square deviation of dihedral angle θ, dihedral angle θ and soluble molecular surface area between backbone atoms of protein gene e and f, which indicates the similarity of the secondary structure of protein gene e and protein gene f (18). Where k, n and t are constants, representing the weight of each parameter.

(C) Then calculate Sim_{CSG_1,CSG_2}:

$$Sim_{CSG_1,CSG_2} = g \times PPA_{CSG_1,CSG_2} + l \times \frac{1}{A_{CSG_1,CSG_2}} \quad (7)$$

Where g and l are constants representing the weight of each parameter.

(D) If the score of the best ratio pair is equal to A_{CSG_1,CSG_2} and there are other corresponding best sets, then step (1)-(3) will be cycled until all the best corresponding sets are used. The minimum value of Sim_{CSG_1,CSG_2}, which is the ASG similarity of the two protein genes, was found.

New standard calculation model

In the last classification result, the ith protein gene parameter with the highest error rate was set as p_i (i.e. RMSD, Z-Score, STR, LOC, ASG), then $\sum p_i$ was the sum of all the protein gene parameters that were wrongly detected, so the mean value of the parameter could be expressed as $\overline{P} = \frac{\sum_{i=1}^{v} p_i}{V}$ then the mean square deviation of the parameter could be expressed as:

$$C = \sqrt{\sum_{i=1}^{v}\left(p_i - \overline{P}\right)^2} \quad (8)$$

Q is the parameter value of the old detection standard of a certain layer of protein gene, then the new detection standard of protein gene Q' can be expressed as:

System operation process

To simulate the operation mechanism of the abnormal data detection system of the protein gene library based on data mining technology, the generation, detection process and evolution of detection agent (protein gene) will be carried out synchronously in the operation process of the system (19). According to the principle of immunity, the working process can be divided into three stages: randomly generating protein gene through gene expression, negative selection, clone selection and copying memory cells (20). The operation process is as follows:

(A) Randomly select a group of protein genes from the protein gene pool for gene expression, which is similar to the genetic gene expression process. A new gene can be generated through the selection, crossover and variation of different gene attributes.

(B) Extract the current data characteristics of the protein gene in the protein gene database (21), which are equivalent to the self-defined dynamically, that is, the normal "self" data characteristics. If the new protein gene data can match any of its data feature patterns, it indicates that the protein gene may mistake the normal protein gene data for the abnormal protein gene data, so delete it and return to step (A), which is a negative selection process (22).

(C) Through the negative selection of mature protein gene grouping, it is transmitted to each host (lymph node) to carry out the actual detection task: through the pre-processing (reduction process), the collected protein gene data characteristics are continuously grouped according to the attack type standard, mapped into the sequence of attribute arrangement, the fuzzy discrete value method and the same coding length as the antibody gene (23), and then it is the same as the antibody gene (23). In the process of gene matching, the second level dynamic optimization algorithm in Section "feature data extraction " is used to calculate the ASG similarity value of protein gene secondary structure arrangement.

(D) If a mature protein gene can match enough groups of protein genes (threshold) in a test cycle, the abnormal protein gene is found If the abnormal protein gene is a known type, start the response program and transfer (6), if it is a new type, send an alarm to the administrator (E); otherwise, it will be deleted as an invalid protein gene, and return to step (A).

(E) When a protein gene is found to be a gene abnormal type, the host sends an alarm to the server by sending a message, and the administrator determines whether it is abnormal data. If the administrator confirms the alarm as abnormal data (24) within a given time, the protein gene will enter the clone selection stage (F); otherwise, if the alarm is considered as a false alarm, the corresponding protein gene will be deleted and returned (A).

(F) Clone selection and replication. Proved effective protein genes are selected and copied to each host node to form memory cells of protein genes. They have a smaller threshold and longer life cycle than common protein genes (25), which can accelerate the detection

process of abnormal data of protein genes previously appeared.

(G) Once a mature protein gene is started or cloned in the test phase, the genes that make up the protein gene will be added to the protein gene library. If the protein gene has been stored in the library, the value representing their fitness will be increased by 1.

(H) In order to avoid the large protein gene pool, a maximum value (determined by the number of type coding digits) can be set in advance. When the length of the protein gene pool exceeds the maximum value, the protein gene with low adaptability will be deleted according to the adaptability value of the protein gene.

Results

System performance test

In order to verify the performance of this system, the experimental results were compared with the improved neural network system, wavelet transform system and fuzzy k-means system. The experimental comparison mainly includes detection rate, false detection rate, system stability and convergence rate.

The selected training set and test set were from the authoritative sequence protein gene library established and maintained by NCBI. There are about 5.94 million kinds of protein genes in this protein gene library after sorting out the duplicate data, among which the abnormal data types of protein genes are mainly divided into four categories and 52 kinds of abnormal data types of protein genes. The gene library contains a labeled training set and unlabeled abnormal data test set. The test data and training data have a different probability distribution. The test data contains some abnormal data types of protein genes that are not in the training set, of which 17 abnormal data types of protein gene only appear in the test set, and the training set contains one normal protein gene data. The identification type and 22 abnormal data types of protein genes make the simulation experiment more practical.

In order to facilitate processing, about 100000 protein genes are selected from the whole gene pool as samples, of which 70000 are selected as training sets, which are used to generate initial autosets, and about 30000 other protein genes are used as test sets. The system in this paper is used to detect the abnormal data of the protein gene. Through the combination of experiment and experience, several important parameters in the experiment are determined: set TCP detector length L=131, UDP detector length L=125, ICMP detector length L=109, matching threshold Sn=0.8, initial self-set=80, immature detector tolerance period ζ=12 generations, mature detector activation threshold β=15, mature detector life cycle λ=6 generations, number of memory detectors R=36, unrecorded The total number of memory detectors M=60, and the total number of iterations N=200. In this paper, through the contrast experiment with the improved neural network system, wavelet transform system and fuzzy K-Means system, through 50 experiments, the average value of the experimental data is obtained, and the contrast figures of detection rate TP and false detection rate FP of four systems are obtained, as shown in Figure 4 and Figure 5 respectively.

It can be seen from Figure 4 that the system in this paper has better detection efficiency. When the iterative algebra of the detector is about 60 generations, the detection performance tends to be stable gradually, and the convergence speed is significantly higher than the improved neural network system, wavelet transform system and fuzzy K-Means system. This is because the system in this paper adopts the combination of random dynamic generation algorithm and multi-point gene recombination algorithm, the immature detector is generated. By using the antibodies of the mature detector and the replaced memory detector, a new effective protein gene is quickly generated by using the multi-site gene recombination rule, which makes the detector easier to pass the negative selection of self-tolerance, increases the conversion rate of the mature detector, shortens the tolerance time, and speeds up the convergence speed of the algorithm. And the updating rules of memory detectors also play an important role in the detection process. By eliminating low-affinity protein genes, the quality of the detector is guaranteed and the detection efficiency is improved. It can also be seen from Figure 4 that the adaptability of the wavelet transform system and the fuzzy K-Means system is relatively poor, and the detection rate has different degrees of vibration, especially the performance of the wavelet transform system, which is caused by the change of the self-set and the non-self-set, and the protein gene in the memory detector has not been updated in real-time. The reason is that we introduce the renewal mechanism of the detector, eliminating the memory detector with low affinity in real-time, ensure the quality of the protein gene in the memory detector, and quickly detect the abnormal data

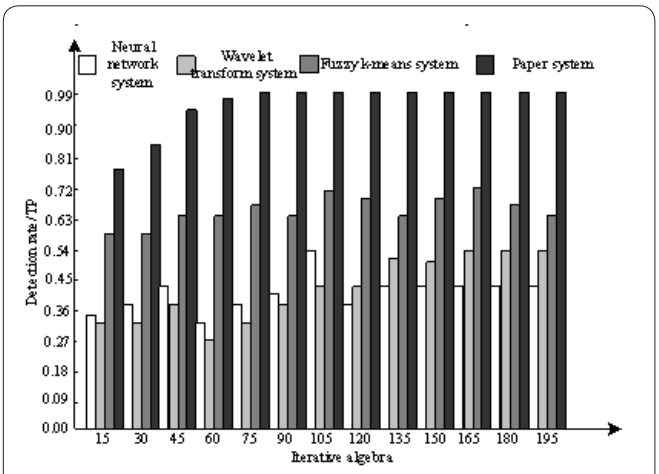

Figure 4. Comparison of detection rate TP.

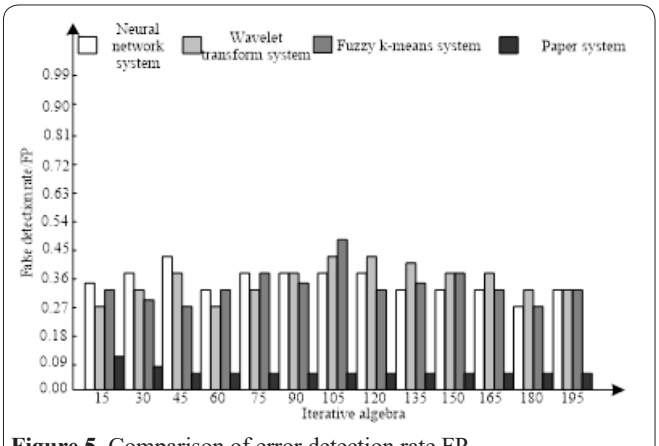

Figure 5. Comparison of error detection rate FP.

of the protein gene library in the current system, so as to ensure the detection efficiency.

It can be seen from Figure 5 that the false detection rate of the detection system in this paper is lower than that of the improved neural network system, wavelet transform system and fuzzy K-Means system. This is because the detector update algorithm in this paper not only solves the pressure caused by the continuous accumulation of the number of detectors in the system but also selectively updates the number and types of detectors in the system through the evaluation of the efficiency of the detector, which not only avoids the reckless deletion of the protein gene of the high-performance memory detector and the decrease of the detection efficiency but also avoids the unreality of the detector. The abnormal data of the protein gene library was misreported due to the updating of the database.

Detection and verification of abnormal data of protein gene

In order to verify that the system can effectively detect the abnormal data in the protein gene database, two sets of protein gene abnormal data are selected, and the two sets of protein gene abnormal data are detected by the improved neural network system, wavelet transform system, fuzzy K-Means system and the system respectively. Each system detects the two sets of protein gene abnormal data five times and compares the two lines The accuracy of the test results is shown in Table 1.

In Table 1, it can be found that when the improved neural network system, wavelet transform system and fuzzy K-Means system are used to detect the abnormal data of two groups of protein genes, the accuracy of five detection results is below 71%. When the system in this paper is used to detect the abnormal data of two groups of protein genes, the accuracy of five detection results of two groups of abnormal data sets of protein genes is the same based on the weight. All of them are between 93.2% and 97.1%, so the detection accuracy of this system is higher. The above experimental results show that the system can improve the accuracy of protein gene abnormal data detection. After more than three months of operation and debugging, it is found that the system has achieved the expected goal. It can automatically and quantitatively detect the abnormal data of the protein gene, and the accuracy of abnormal data detection is over 97.1%.

In order to verify that the system in this paper can effectively detect the abnormal data of protein gene in the protein gene library, the improved neural network system, wavelet transform system, fuzzy K-Means system and the system in this paper are used to detect the abnormal data of 5.94 million protein genes in the protein gene library, and the detection results are shown in Table 2.

Table 2 shows that when using this system to detect the abnormal data of protein genes in the protein gene library, 5.73 million protein genes were detected as normal protein genes, 100000 protein genes were detected as abnormal data protein genes and 110000

Table 1. Accuracy comparison of four experimental data sets in three tests.

Neural network system							**Wavelet transform system**						
Testing		**For the first time**	**Secondary**	**The three-time**	**The fourth time**	**The fifth time**	**Testing**		**For the first time**	**Secondary**	**The three-time**	**The fourth time**	**The fifth time**
	k	1	1	1	1	1		k	1	1	1	1	1
	n	1	1	1	1	1		n	1	1	1	1	1
Weights	t	1	1	1	1	1	Weights	t	1	1	1	1	1
	g	1	1	2	2	0.001		g	1	1	2	2	0.001
	l	1	1	1	2.5	0.001		l	1	1	1	2.5	0.001
Accuracy rate	H1	54.30%	53.20%	52.50%	61.60%	62.90%	Accuracy rate	H1	57.30%	61.20%	66.50%	64.60%	62.90%
	H2	54.80%	56.40%	55.70%	60.20%	64.30%		H3	59.60%	58.00%	66.30%	62.40%	61.70%
Fuzzy k-means system							**Paper system**						
Testing		**For the first time**	**Secondary**	**The three-time**	**The fourth time**	**The fifth time**	**Testing**		**For the first time**	**Secondary**	**The three-time**	**The fourth time**	**The fifth time**
	k	1	1	1	1	1		k	1	1	1	1	1
	n	1	1	1	1	1		n	1	1	1	1	1
Weights	t	1	1	1	1	1	Weights	t	1	1	1	1	1
	g	1	1	2	2	0.001		g	1	1	2	2	0.001
	l	1	1	1	2.5	0.001		l	1	1	1	2.5	0.001
Accuracy rate	H1	66.10%	64.20%	65.70%	62.70%	61.90%	Accuracy rate	H1	94.30%	93.20%	95.50%	96.60%	96.90%
	H3	70.50%	69.40%	67.10%	64.80%	68.40%		H3	95.60%	95.00%	96.30%	94.30%	97.10%

Table 2. Comparison of protein gene detection results in different systems.

Naive Bayesian detection system			**Paper system**			**Naive Bayesian detection system**			**Paper system**		
Normal data of protein gene (Ten thousand)	Abnormal data of protein gene (Ten thousand)	Doubt (Ten thousand)	Normal data of protein gene (Ten thousand)	Abnormal data of protein gene (Ten thousand)	Doubt (Ten thousand)	Normal data of protein gene (Ten thousand)	Abnormal data of protein gene (Ten thousand)	Doubt (Ten thousand)	Normal data of protein gene (Ten thousand)	Abnormal data of protein gene (Ten thousand)	Doubt (Ten thousand)
555	39	0	541	53	0	549	45	0	573	10	11

protein genes were detected as suspected abnormal data protein genes. Based on the same situation, when using the improved neural network system, wavelet transform system and fuzzy K-Means system to detect the abnormal data of protein gene pool, 5.55 million, 5.41 million and 5.49 million protein genes were detected as normal protein genes, and 390000, 530000 and 450000 protein genes were detected as abnormal data protein genes. Using this system, 100000 normal protein genes are detected as abnormal protein genes, while using other detection systems, the detected abnormal data of protein genes are much higher than that of this system. It can be seen that, in the same case, using this system can greatly reduce the error rate of detecting normal protein genes as abnormal protein gene, so this system can effectively detect abnormal protein gene data, and the detection accuracy is very high.

Discussion

Because there are many kinds of protein in the protein gene library, and the quantity of protein is huge, there will be the abnormal situation of protein gene in the huge quantity of protein gene library, and the abnormal situation of protein gene will cause the development abnormality, disease and even death of animals and plants, so it is very necessary to detect the abnormal data of protein gene. This paper studies the abnormal data detection system of the protein gene library based on data mining technology. The innovation of this paper lies in the ingenious introduction of data mining technology into the abnormal data detection of the protein gene library, and the formation of a detection system (26-29). This paper verified the effectiveness of this system from two aspects of system performance and abnormal data detection of the protein gene.

(A) System performance verification: the experimental results of this system are compared with the improved neural network system, wavelet transform system and fuzzy K-Means system, it is found that this system has a better detection rate, and the convergence speed is significantly higher than the improved neural network system, wavelet transform system and fuzzy K-Means system. This system is used to detect the abnormal data of the protein gene library, and the detection rate is stable. In this paper, the false detection rate of the detection system is lower than the improved neural network system, wavelet transform system and fuzzy K-Means system. Therefore, it can be seen that this system has a high detection rate, good convergence and low false detection rate when detecting the abnormal data of the protein gene library. Therefore, the abnormal detection of the nucleotide gene library and DNA database in the biological field can be realized by this system. This system has a broad application prospect in the future detection of abnormal data of biological gene library.

(B) Detection of abnormal data of protein gene: in order to verify that the system can effectively detect the abnormal data in the protein gene library, the experimental results of the system are given. By using the improved neural network system, wavelet transform system, fuzzy K-Means system and this system to detect the experimental data set of two groups of protein genes five times, and comparing the accuracy of the four systems, it is found that this system improves the detection accuracy of abnormal data of protein genes, and this system can automatically and quantitatively classify protein gene structure The accuracy is over 97.1%.

In order to verify that this system can effectively detect the abnormal data of protein genes in the protein gene library, we also use the above four systems to detect the abnormal data of 5.94 million protein genes in the protein gene library. It is found that this system can greatly reduce the error rate of detecting normal protein genes as abnormal protein genes, so this system can effectively detect protein genes abnormal data and high detection accuracy.

In the operation, we also found that there are two places to continue to improve: at present, there are nearly 20 million proteins tested by this system, but the amount of data is far from enough to develop more accurate detection standards. Secondly, the new detection standard is based on the detection parameters with the highest error rate. In fact, the trend of the accuracy rate is not always increasing, and no further optimization method is given in this paper. When the accuracy rate of detection decreases, a simple way can be used, that is, when the accuracy rate shows a downward trend, stop using the new detection standard, and use the detection standard that brings the highest accuracy rate before.

In this system, the firewall module is introduced to filter the protein gene entering the system, which effectively reduces the pressure of system monitoring and detection; immature detector module, mature detector module and memory detector module in immune module promote the detection of abnormal protein gene data according to their characteristics. The first randomly generated detector module will initialize the immature detector according to the AUTOSET, and carry out a variation on the extracted protein gene sequence, generate new protein gene sequence, and complete the diversity transformation of protein gene; the integrity of the mature detector directly affects the accuracy of the system in detecting the abnormal data of protein gene; the memory detector in detecting the abnormal data of protein gene library. When the system is attacked by the same or similar abnormal protein gene data again, it can react quickly.

At the same time, the data mining technology is introduced into the abnormal data detection of protein gene in this system. The similarity value of two protein genes is obtained by calculating AGS, and the abnormal gene features in the protein gene library are extracted. Then the protein gene is randomly generated by gene expression, and the abnormal data detection of the protein gene library is realized by the immune principle. Experiments show that the system can effectively improve the detection rate and reduce the false detection rate. The system has good dynamic real-time detection ability in protein gene detection.

References

1. Jiu PX, Kai S, Lei X. Integrated system health management-oriented maintenance decision-making for multi-state system based on data mining. Int J Syst Sci 2016; 47: 15-23

2. Yan HY, Yu XZ, Yang G. Analysis of the autophagy gene expression profile of pancreatic cancer based on autophagy-related protein

microtubule-associated protein 1A/1B-light chain 3. World J Gastroenterol 2019; 25: 2086-2098.
3. Ernur S, Benjamin J, Harrison KW. Framework for reanalysis of publicly available Affymetrix® GeneChip® data sets based on functional regions of interest. BMC Genomics 2017; 18: 875-889.
4. Ming Z, Gerold SU, Christine S. Drug repositioning for Alzheimer's disease based on systematic 'omics' data mining. Plos One, 2016; 11: 168-181.
5. Shon HS, Han SH, Kim KA. Proposal reviewer recommendation system based on big data for a national research management institute. J Inform Sci 2016; 43: 147-158.
6. Zhe HZ, Zhi HY, Hong L. A protein-protein interaction extraction approach based on deep neural network. Int J Data Mining and Bioinformatic 2016; 15: 145-167.
7. Jia HB, Yi FT, Zhe WQ. ClickGene: An open cloud-based platform for big pan-cancer data genome-wide association study, visualization and exploration. BioData Mining 2019; 12: 114-152.
8. Zhang YQ, Wang Y, Liu HD, Li B. Six genes as potential diagnosis and prognosis biomarkers for hepatocellular carcinoma through data mining. J Cell Physiol 2018; 234: 332-339.
9. Marina M, Dora I, Ríos C. Determining clostridium difficile intra-taxa diversity by mining multilocus sequence typing databases. BMC Microbiol 2017; 17: 62-71.
10. Rajalkshmi D, Dinakaran K. A novel time series pattern matching model combined with ant colony optimization and optimal binary search trees based segmentation approach. J Comput Theor Nanosci 2017; 14: 5203-5208.
11. Hong BX, Hai XW, Ming ZQ. In silico drug repositioning for the treatment of Alzheimer's disease using molecular docking and gene expression data. RSC Adv 2016; 6: 98080-98090.
12. Jaewon C, Hyuk JK. The information filtering of gene network for chronic diseases: Social network perspective. Int J Distrib Sens Netw 2015; 20: 1-6.
13. Jin WB, Rosa Y, Kenneth DC. Fungal artificial chromosomes for mining of the fungal secondary metabolome. BMC Genomics 2015; 16: 343-361.
14. Zhao XW, Fei F, Xiao SZ. Development of diagnostic model of lung cancer based on multiple tumor markers and data mining. Oncotarget 2017; 8: 94793-94804.
15. Song JD, Simon XY, Feng CT. Classification of orange growing locations based on the near-infrared spectroscopy using data mining. Intell Autom Soft Comput 2015; 22: 1-7.
16. Devi F, Achmad NH, Jonson LG. A spatio-temporal data-mining approach for identification of potential fishing zones based on oceanographic characteristics in the eastern Indian Ocean. IEEE J Sel Top Appl Earth Obs Remote Sens 2015; 9: 1-9.
17. Fereshteh C, Mehdi S, Elahe E. Confident gene activity prediction based on single histone modification H2BK5ac in human cell lines. BMC Bioinformatics 2017;18: 67-89.
18. Hongya Z, Debby DW, Long C. Identifying multi-dimensional co-clusters in tensors based on hyperplane detection in singular vector spaces. Plos One 2016; 11; 162-293.
19. Hongyan L, Xi XM, Chun XW. Design and analysis of a general data evaluation system based on social networks. EURASIP J Wirel Commun Netw 2018; 218: 109-127.
20. Yu XL, Xu YJ, Zhou ZX. Sparse event detection based on parallel discrete social spider optimization algorithm and compressed sensing in wireless sensor networks. J Chin Acad Electron Inform Technol 2017; 12: 202-208.
21. Zhang CH, Zhou JW, Du CS. Review of control strategies of single-phase cascaded h-bridge multilevel inverter for grid-connected photovoltaic systems. J Power Supply 2017; 15: 1-8.
22. Hu NJ, Zhou W, Zheng JL. Preparation and electrochemical performance of porous V2O5 microspheres. Chin J Power Sources 2018; 42: 108-116.
23. Qu JJ. Research on the function of electronic medical record and related problems in hospital informatization management. Autom Instrum 2017; 15: 226-227.
24. Tang M, Yang Y, Li XF. Two-grid finite element discretization methods for a class of Poisson-Nernst-Planck equations. J Jilin Univ (Science Edition), 2019; 57: 71-77.
25. Liu W, Xu CH, Chen ZY. Simulation of adaptive filtering method for target image data. Comput Simul 2017; 34: 260-263.
26. Talat F, Wang K. Comparative Bioinformatics Analysis of the Chloroplast Genomes of a Wild Diploid Gossypium and Two Cultivated Allotetraploid Species. Iran J Biotechnol 2015; 13(3): 47-56.
27. Son J, Jeong H, Lee E, No S, Park D, Chung H. Identification of specific gene expression after exposure to low dose ionizing radiation revealed through integrative analysis of cDNA microarray data and the interactome. Int J Radiat Res 2019; 17 (1) :15-23.
28. Kuai H, Zhong N. The Extensible Data-Brain Model: Architecture, Applications and Directions. J Comput Sci 2020: 101103.
29. Dash S, Shakyawar SK, Sharma M, Kaushik S. Big data in healthcare: management, analysis and future prospects. J Big Data 2019; 6(1):54.