Emerging Patterns and Classification Algorithms for DNA Sequence

Xiaoyun Chen
College of mathematics and computer science , Fuzhou University,Fuzhou,China
Email: c_xiaoyun@21cn.com

Jinhua Chen
College of mathematics and computer science , Fuzhou University,Fuzhou,China

***Abstract*—Existing machine learning methods for classification of DNA sequence achieve good results, but these methods try to express a DNA sequences as discrete multi-dimensional vector, so when the length of the sequences in the DNA sequence database is not fixed or there exists some omitted characters, these methods can not be used directly. In this paper, we define the new support and growth rate of support to find the frequent emerging patterns from DNA sequence database, and present a classification algorithm FESP based on the frequent emerging sequence patterns. The frequent emerging sequence patterns keep the information provided by the order of bases in gene sequences and can catch interaction among bases. FESP algorithm applies classification rules that are constructed by frequent emerging sequence patterns of each class to classify the new DNA sequences. This method can work on sequences with different lengths or omitted character and shows good performance.**

***Index Terms*—emerging sequence pattern, classification rule, feature selection, DNA**

I. Introduction

As the Human Genome Project started in 1990, DNA sequence data containing hundreds of millions of base pairs and tens of thousands of protein structures have been determined. The generation of large amount of biological sequence data promoted the birth of bioinformatics, which makes people seek various kinds of ways and means only to obtain knowledge they need from the massive biological sequence data and to reveal the nature of various biological phenomena. Study of classification of biological sequences has been developed for more than 10 years and has achieved some progress. Especially, the introduction of Support Vector Machine (SVM) these years greatly improves the classification precision [1]. But most of the existing classification methods including support vector machines express DNA sequences as some discrete value or multi-dimensional vectors, ignore the order of the sequences and interaction among amino acids. In addition, when length of sequences in the DNA sequence database is not fixed or there exists some base omitted, these methods based on multi-dimensional vectors can not be used directly.

Association rule mining is to find all rules that meet specified minimum support threshold and minimum confidence threshold by thorough searching. During association rule mining, computational complexity and the number of rules grow up exponentially, and the minimum support become the key to control the growth. As research on the association rules being further, the researching focus has turned from improving the efficiency of association rule mining to using association rules for practical problems. Here, Classification based on association rules is a useful attempt, and CBA [2] and CMAR [3] are representatives of such algorithms. A common characteristic of these methods is using frequent patterns to express the relation among attributes or features of samples. If some attributes or attribute set of some samples appear frequently, we think the attributes or attribute set can be used to describe the common character of these samples, that is, there is a strong correlation between the attributes or attribute set and these samples. In 1999, Dong and Li present a concept of Emerging Pattern in paper [4]. Different from frequent pattern with the support greater than a minimum threshold, the support of emerging pattern increases significantly from one class to another class. Hereafter, in paper [7] they present a genetic data classification method using a particular jumping pattern only appearing in one class. In this paper, we propose a new emerging pattern based gene data classification method. This method try to choose emerging patterns involved in classification according to the differentia in frequency of patterns appearing in different classes and then construct classification rules. Compared with Support Vector Machine (SVM) and Neural Network, emerging pattern based classification rules can catch interaction among bases and provide a clear description for causal

Manuscript received October 1, 2010; revised January 15, 2011; accepted January 20, 2011.

doi:10.4304/jsw.6.6.985-992

relationship between bases and functions, so it is easy to understand.

II. Feature selection

Gene sequence data set is a mixed set consisting of signals from various biological senses, so there is in a mass of noise. Moreover, high dimensional feature lead to the case that there are only small part of fragments in large number of gene sequences related to the category of samples. Therefore, we have to do feature selection before the implementation of the classification algorithm on the gene sequence data in order to pick up fragments most related to classification and reduce the dimension of feature gene, which may lead to more concise classification rules and improve the efficiency of classification.

Problem of feature selection can be viewed as an optimization problem. Its searching space is all possible feature subsets, which means the scale of its searching space is

$$\sum_{k=0}^{m}\binom{m}{k}=2^{m} \qquad （1）$$

where m is the number of features in original feature set and k is the scale of feature set to be selected.

For high-dimensional feature spaces of gene sequences, many feature selection methods of machine learning are no longer applicable. Then, the assumption of feature independence is usually used to simplify the problem, aiming at save some time by giving up a little quality. Therefore, feature selection methods used for high-dimensional gene sequences are quite easy, compared to machine learning methods.

The general feature selection methods for gene sequence aim at evaluate each original feature using evaluation functions. Each one of the features is evaluated separately to calculate the score. The steps of feature selection based on evaluation functions can be summarized as follows:

1) Initially, the feature set contains all original features.

2) To calculate the value of evaluation function for each feature in the feature set.

3) To sort the values of evaluation functions for all features.

4) To choose the top k features (k is the number of features we need) as the feature subset. There is not a good solution to ascertain the specific number of features to be selected. We can give an initial value and then gradually adjust it along with the experimental testing and according to the statistical results to find the best value.

For gene sequences divided into different classes, the feature evaluation function can refer to the existing methods in text feature selection, such as Information Gain (IG), Cross Entropy (CE), Mutual Information (MI) and χ^2 statistic [7]. Because χ^2 statistic can is used for the representation of correlation between feature variables and class variables, we adopt Pearson's χ^2 statistic as the feature evaluation function.

Let D be a set of DNA sequences and it contains c classes, C_1, C_2, …, C_c. The number of samples of ith class is N_i, and there are totally N samples. Each DNA sequence is a symbol sequence composed of four characters A, T, C and G, denoted as $Y_m =< x_1x_2…x_m >$, $x_i\in\{$A,T, C, G$\}(i=1,2,..,m)$.

Feature x of DNA sequences may be A, T, C or G, separately denoted as A_1, A_2, A_3 and A_4. And then the χ^2 statistic of x can be calculated by the following function [7]:

$$\chi^2(x,C_j)=\sum_{i=1}^{4}\frac{(O_{ij}-E_{ij})^2}{E_{ij}} \qquad （2）$$

here O_{ij} is the number of samples which are in C_j class and whose feature x equal to A_i; R_i is the number of samples whose feature x equal to A_i in D; N_j is the number of samples in C_j class and $E_{ij}=R_i*N_j/N$.

The value produced by χ^2 statistic function describes the importance degree of the feature to specifical class. To evaluate the importance of a feature to the whole sample set, we have to synthetically consider the importance of the feature to each class and assign the global score to be the average value of the χ^2 statistic function of the feature to each class. The specific formula is as follows:

$$\chi^2_{avg}(f)=\sum_{j=1}^{c}\frac{N_j}{N}\chi^2(f,C_j) \qquad （3）$$

III. Classification based on frequent emerging sequence patterns

The current sequence pattern mining research is carried out from two aspects: First, find the repeat patterns in a single sequence. In sequence mining, the repeat pattern is the continuous sub-sequence that occurs frequently in a sequence. As the needs of evolution, biology may own a large number of replication sequence fragments that are replicated by themselves. Perhaps the functions of these repeats fragments are unknown, but all have important biological significance. Second, mining the conservative frequent patterns in a biological sequence set that is composed of multiple biological sequences. Although nature has undergone millions of years of biological evolution, but some fragments in biological sequences play a more important role for biological survival, thus showing relative stability during the evolution, which is conservative. Such as some conservative sequence patterns of protein family plays a key role for the protein structure and function.

Frequent Emerging Sequence Patterns based classification algorithm (FESP) proposed in this paper is intent to classify the DNA sequences in DNA sequences database using the emerging patterns based associative classification frame. For the particularity of DNA sequences data, the existing associative classification algorithms, such as CBA [2], CMAR [3], ARC-BC [5] and CAEP [6], can not be used directly. We have to give some necessary definitions and do some transformations. The two main transformations are as follows:

(1) Different from the traditional definition of frequent itemsets and emerging patterns, taking into account that four characters ATCG of DNA sequences are arranged in order, the adopted emerging sequence patterns can keep the order of the items instead of emerging patterns.

(2) In order to reduce the number of classification rules, we choose the emerging sequence patterns which have stronger classification ability to construct classification rules that is adopting the frequent emerging sequence patterns, which is frequent in a single sequence and the support growth of which in the whole data set is greater than the given threshold.

Like the classical classification algorithms, frequent emerging sequence patterns based classification algorithm also includes two stages of training and classification. In the training stage, the frequent emerging sequence patterns are mined according to the user-specified constraints and then use the frequent emerging sequence patterns to construct the classification rule set. In the classification stage, we use the classification rule set obtained in the training stage to classify the new DNA sequences.

A. Basic Terms and related research

TABLE I. BASIC TERMS

1.Alphabet $\sum$: the characters set appearing in DNA sequence set D, $\sum$={A,C,G,T}

2.Sequence Y_m: Y_m=<$x_1x_2\ldots x_m$>, where $x_i\in\sum$, i=1,2,…,m, m \denotes the length of sequence Y_m.

3. Subsequence y_k: Subsequence y_k=<$x_Lx_{L+1}\ldots x_{L+k}$> is a continuous segment starting in the ith characters position of the sequence in sequence set D, here k denotes the length of subsequence y_k.

Definition 1 Classification rule

Assume y_k is a particular subsequence of DNA sequences set D, C={C_1,C_2,…,C_c} is the set of category label, the classification rule is an implication as following form:

$$y_k \Rightarrow C_i \qquad (4)$$

here y_k is called rule antecedent, C_i is called rule consequent, and the class C_i indicated by the consequent of rule is called the target class.

The support and confidence measure of classification rule based on frequent itemsets must greater than the given minimum support threshold and minimum confidence threshold, respectively.

According to the different engendering methods of the frequent itemsets used to construct the classification rules, association classification can be divided into two categories: global association classification and local association classification [4]. Global association classification (as shown in Fig.1) obtains frequent itemsets from whole training set, and the class label becomes an item of frequent itemset. Local association classification (as shown in Fig. 2) mines the frequent itemsets in each class and constructs the classification rules through these frequent itemsets. This process is so simple that we only need to consider the frequent itemsets of each class as the antecedent and consider the class label as the consequent.

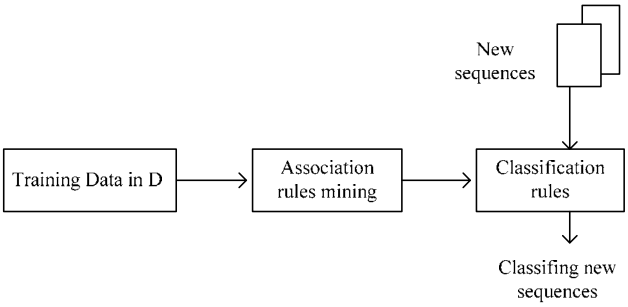


Figure 1. Global association classification.

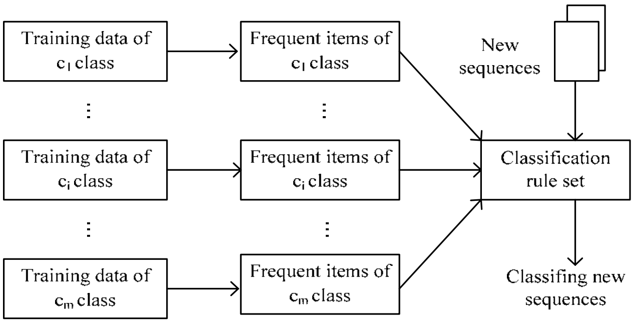

Training data of c1 class
Frequent items of c1 class
New sequences
Training data of ci class
Frequent items of ci class
Classification rule set
Training data of cm class
Frequent items of cm class
Classifing new sequences

Figure 2. Local association classification

B. FESP algorithm

We adopt the local method (as shown in Fig. 2) to building the classification rules. Frequent Emerging Sequence Patterns based Classification Algorithm (FESP) proposed in this paper is intent to classify the DNA sequences in DNA sequences database on the association classification frame. Of course, for the particularity of DNA sequences data, the existing association classification algorithms, such as CBA [2], CMAR [3], ARC-BC[5] and CAEP [6], can not be used directly. We have to give some necessary definitions and do some transformations. The two main transformations are as follows:

(i) Different from the traditional definition of frequent itemsets, taking into account that four characters ATCG of DNA sequences are arranged in order, the adopted emerging sequence patterns to building classification rules can keep the order of the characters instead of disorderly emerging patterns.

(ii) In order to reduce the number of classification rules, choose the frequent emerging sequence patterns that have stronger classification ability to construct classification rules. The frequent emerging sequence pattern is frequent in a single sequence and its support growth is greater than the given threshold among the different classes.

Like the classical classification algorithms, the Frequent Emerging Sequence Patterns based Classification Algorithm also includes two stages of training and classification. In the training stage, we mine frequent emerging sequence patterns according to the user-specified constraints and then use the frequent emerging sequence patterns to construct the classification rule set. In the classification stage, we use the classification rule set obtained in the training stage to classify the new DNA sequences.

1) *Training phase*

Generally speaking, the frequent sequence patterns contain three types: The first is the repeating patterns in biological sequences. Such sequence pattern frequently appears in the same sequence. As the needs of biological evolution, a lot of repeating patterns are produced through the self-replicating of subsequence fragment; The second is the sequence patterns which frequently appear in a sequences set. The emergence of such patterns is due to existence of conserved segments in biological sequences, which has a crucial role for biological survival and shows the relatively stable characteristics during evolution. The third is the sequence patterns which frequently appear both in a single sequence and in several sequences of a sequence set. Xiong and Zhu[8] pointed out that the three patterns are important, any individual pattern may contain only incomplete biological information.

These three definitions of frequent sequence patterns are all based on the distribution of sequence patterns in a single sequence or in several sequences of a sequence set, without consideration of the different distribution of sequence patterns in different sequence sets. For different frequent sequence patterns having different classification ability, classification ability of subsequences only appearing in one class is obviously greater than that of subsequences appearing in all classes. Therefore, we have to choose subsequences with greater classification ability in all frequent subsequences, which we call frequent emerging sequence patterns. The definition of frequent sequence patterns in a single sequence is as follows.

Definition 2 Local Support[8]

Subsequence occurrence frequency in any one of sequences in D is called local support. i.e. given a DNA sequence Y_m with length m, the local support of its continuous subsequence y_k with length k is defined as

$$S(y_k \mid Y_m) = P(y_k \mid Y_m) = \frac{count(y_k \mid Y_m)}{m-k+1} \qquad (5)$$

here $count(y_k|Y_m)$ denote the frequence of the sub sequence y_k appearing in the sequence Y_m. If the support $S(y_k|Y_m)$ is greater than the given minimum support threshold min_S, y_k is a frequent sequence pattern. Note that the length of each sequence is different.

Example 1 Given the support threshold 0.2, the sub sequence y_3= {TGC} in the DNA sequence Y_{10}= {AATT GCTTGC} is a frequent sequence pattern with length 3.

For the DNA sequence set $D=D_1 \cup D_2 \ldots \cup D_c$ containing c classes, where D_i is the sequence set of class C_i, D_i includes N_i sequences and the sequence set D totally includes N sequences. We use the growth rate [7] to describe the difference of support for y_k in different classes.

Definition 3 Total Support

Subsequence occurrence frequency in all sequences in D_i is called total support. i.e. the total support for sequence Y_k in sequence set D_i is defined as

$$S(y_k \mid D_i) = \frac{N_i}{N} P(y_k \mid D_i) \qquad (6)$$

here $P(y_k \mid D_i)$ denote the probability of the sequence pattern y_k in class C_i.

Definition 4 Growth rate of support

Let y_k be a frequent sequence pattern that local support is greater than the support threshold min_S, and the total support of y_k in class C_i is $S(y_k \mid D_i)$, and then the growth rate of support for y_k in different classes is defined as follows:

$$J(y_k \mid D_i) = \sum_{i \neq j; j=1}^{c} [\frac{S(y_k \mid D_i)}{S(y_k \mid D_j) + 0.0001}] P(y_k) \qquad (7)$$

here P(y_k) is the ratio of the number of DNA sequences containing y_k and the number of DNA sequences of the whole sequence set. The 0.0001 in the denominator intends to avoid overflowing when $S(y_k|D_j)$=0.

For example, if $S(y|D_1)$=0.9, $S(y|D_2)$=0.3, $S(y|D_3)$=0.5 and $P(y)$=0.3, then $S(y|D_1)/(S(y|D_2)+0.0001)$=3, $S(y|D_1)/(S(y|D_3)+0.0001)$=1.8, so we have $J(y|D_1)$=(3+1.8)*0.3 =1.44. We can see from the definition 4, the greater the growth rate for y_k in class C_i is, the greater the difference between the support of y_k in class C_i and the support of y_k in other classes will be, and thus the stronger the recognition ability of y_k for class C_i is. Therefore, the pattern y_k is appropriate to construct classification rules.

We name frequent sequence patterns with the top k greatest growth rate in class C_i as frequent emerging sequence patterns (FESP). Because the frequent emerging sequence pattern is a DNA sequence segment that meets a certain threshold, and is important to reflect the structure of DNA sequences, it can be used to express the characteristics of sequence or identify the new sequence. The more a sequence contains the frequent emerging sequence patters in a category, the greater the likelihood it belongs to this category. Therefore, the classification rules $y_k \Rightarrow C_i$ are constructed by frequent emerging sequence pattern y_k as the antecedent and the class label C_i of y_k as the consequent.

Definition 5 Rule confidence

Let y_k be a frequent emerging sequence pattern in class C_i, the confidence of classification rule $y_k \Rightarrow C_i$ is

$$Conf(y_k \Rightarrow C_i) = S(y_k \mid D_i) / p(y_k) \qquad (8)$$

The greater the confidence of classification rule $y_k \Rightarrow C_i$ is, the greater the degree of pattern y_k appearing in class C_i is and the more possible DNA sequences containing y_k belonging to class C_i is. If the confidence is 100%, pattern y_k appears only in class C_i and its classification ability is the strongest.

2) *Classification phase*

Definition 6 coverage rules

If the antecedent of rule r: $y_k \Rightarrow C_i$ is the subsequence of sequence Y, then the rule r is called as coverage rule of sequence Y or the sequence Y is covered by rule r.

For any one of the sequences to be classified, there may be more than one coverage rules pointing to different target classes. Then how can we classify them according to these coverage rules? The association classification algorithm CBA adopts the coverage rule with the highest priority to classify new samples to be classified, completely ignoring other coverage rules of them. ARC considers all the coverage rules of the samples to classify. Firstly, the sum of confidence with the same target class is computed and the label of the new sample is assigned to the target class which has the greatest confidence sum [3].

Definition 7 Class confidence

Probability of sequence Y_m belonging to class C_i is the confidence sum of coverage rules of class C_i. We define the confidence sum as class confidence of Y_m belonging to class C_i that is

$$\Omega(Y_m, C_i) = \sum_{\substack{\mathrm{R_i\ cover\ Y_m} \\ \mathrm{i=1}}}^{l} conf(R_i) \qquad (9)$$

where l is the number of rules covering sequence Y_m and R_i is the rules set covering sequence Y_m in class C_i.

TABLE II. FREQUENT EMERGING SEQUENCE PATTERNS BASED CLASSIFICATION ALGORITHM

Algorithm FESP

Training phase

(1) Find the frequent sequence patterns in each classes of data set D, that is to find sequence patterns with the support greater than the minimum support threshold min_S in a sequence.

(2) Compute the growth rate J of support for all frequent sequence patterns of each classes C_i (i=1,2,…,c), sort them by J and choose the frequent sequence patterns with the top k greatest J in class C_i as frequent emerging sequence patterns (FESP).

(3) Use the frequent emerging sequence patterns and the respective class labels C_i to construct classification rule like $y_k \Rightarrow C_i$, and compute the rule confidence

$$Conf(y_k \Rightarrow C_i) = S(y_k \mid D_i) / p(y_k)$$

Classification phase

(4) Find all coverage rules of Y in the classification rule set.

(5) Coverage rules are grouped by class labels and rules in the same group have the same class label. Compute the confidence sum Ω of rules of each group that is class confidence.

(6) Y is assigned into the class with the greatest class confidence Ω.

$$C_k = \underset{i}{Max}\arg\Big(\sum_{\substack{y_j \in C_i \\ y_j \in Y}} Conf(y_j \Rightarrow C_i)\Big)$$

Ⅳ EXPERIMENTAL RESULTS

We adopt the HS3D (Homo Sapiens Splice Sites Dataset) [9] as the experimental data set, which includes 2231 sequences together and is divided into 3 groups by the positions of DNA sequence fragments, as follows:

The 1st group: intron exon, marked as IE

The 2nd group: exon intron, marked as EI

The 3rd group: neither IE nor EI, short for N

After wiping off the sequences containing non-base symbols N, D and R from the original sequence set, EI includes 762 DNA sequences, IE includes 765 and N includes 704. We select the first 200 sequences of IE and EI and select the first 600 sequences of N to be the training samples. The rest sequences are testing samples. Thus the 1000 training samples are 200 (IE), 200 (EI), 600 (N). The 1231 testing samples are 562 (IE), 565 (EI), 104 (N). The number of features of the original HS3D data set is 60.

A. Experiments of FESP

Fig.3 shows the number of frequent sequence patterns based on different minimum support thresholds. We may find that with the minimum support threshold increases, the number of frequent sequence patterns becomes less obviously.

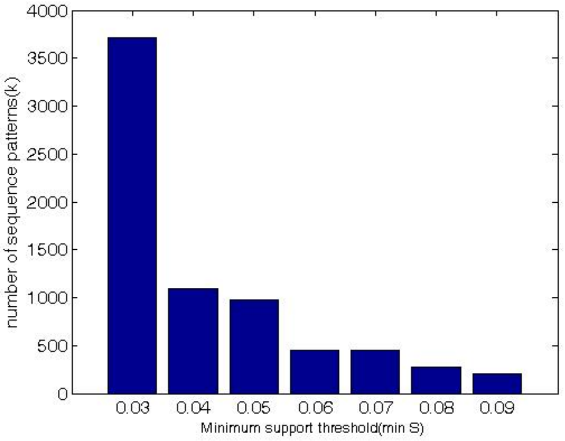


Figure 3. Number of frequent sequence patterns with different support threshold

Fig.4 shows the classification results of FESP based on different minimum support thresholds. Here k=all means that we choose all frequent sequence patterns to construct classification rules, while k=the best mean to achieve the highest accuracy with the best k, such as the best k is 300 and 60 respectively when min_S=0.03 and min_S=0.04 , that means to adopt frequent sequence patterns with top 300 or 60 greatest growth rate to construct classification rules. Through comparing the left and right diagrams of Fig.4, we can see that it doesn't mean the more the frequent sequence patterns used for constructing classification rules are, the better the results are. Only selecting some of the frequent sequence patterns with greater growth rates shows better performance than choosing all frequent sequence patterns to construct classification rules, and it results in higher classification precision, its accuracy has an improvement of more than 30%. The right diagram of Fig.4 shows the best results of different k on different minimum support threshold. We may find that as the minimum support threshold becomes lower, the classification accuracy of FESP obviously increases. Fig.5 shows classification accuracy using different number of emerging patterns to construct classification rules when the minimum support threshold is 0.03. From Fig.5, we found that the classification accuracy increases with the number of emerging patterns increasing at first, but when classification accuracy reach a maximum 0.645004（when k=300）, it begins to fall. It is similar at other minimum support thresholds. This shows that using only a small amount of sequence patterns with the highest growth rate to construct classification rules may not be necessarily to get the best results. It is because too few patterns lead to too few classification rules produced, which makes sequences to be classified not be covered by any rules and thus can not be classified. Instead, if we choose too many sequence patterns, some sequence patterns with low growth rate (even 0) may be used for constructing classification rules, which interferes the evaluation of classification. That is to say, some sequence patterns with low growth rate have great rule confidences which play a major role in classification.

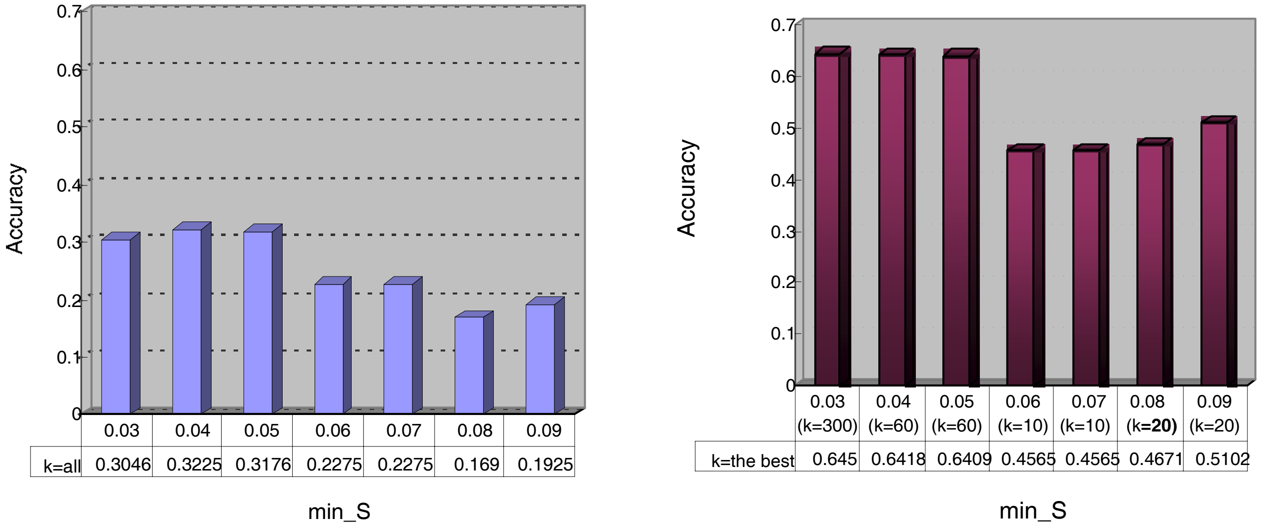


Figure 4. Classification result with different minimum support thresholds

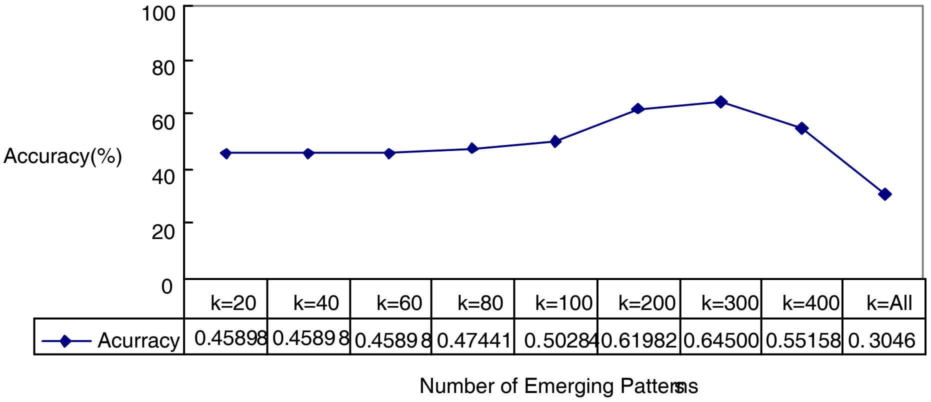


Figure 5. Effect of number of emerging patterns on classification accuracy

B. Experiments of FESP algorithm based on feature selection

If we consider each base of DNA sequences in HS3D as a feature, then we can use χ^2 statistic to measure all features in HS3D and sort the scores descending, and consequently the continuous features with the highest scores can be regarded as the sequence fragments with the strongest classification ability in the data set. For the data set HS3D, we get 8 continuous features with the highest scores, from the 28th to the 35th features. By selecting the 28th to the 35th bases of all sequences in HS3D training set and testing set, we obtain 8-lengthed training subset with 1000 samples and testing subset with 1231 samples.

Table 3 shows the classification results for the subset of HS3D. Comparing Fig.3 and Fig.4, we may find that after feature selection, the classification accuracy is obviously improved although length of sequences was reduced to 8. This also shows that in every DNA sequence, different sequence fragments contain different information, so by sequence fragments obtained by feature selection we can not only improve the classification ability but also reduce the scale of data sets and thus improve classification efficiency.

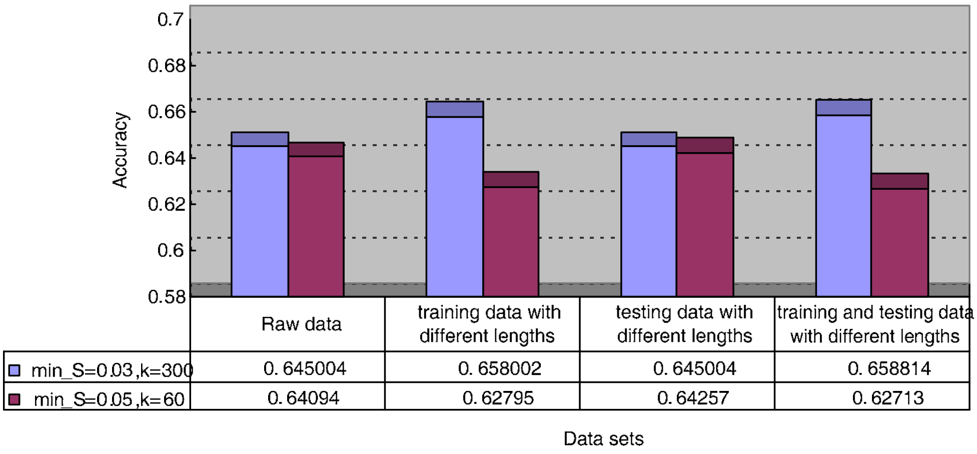
TABLE III. CLASSIFICATION RESULTS BASED ON FEATURE SELECTION (8 FEATURES)

min_S	0.03~0.9	0.1	0.2	0.25
number of sequence patterns	6198	6198	2012	537
Accuracy	0.495532 (k=all)	0.495532 (k=all)	0.708367 (k=all)	0.709992 (k=all)
	0.689683 (k=300)	0.689683 (k=300)	0.708367 (k=300~1000)	0.709992 (k=300~537)

C. Classification experiments of FESP on sequences with different lengths

In order to test the classification results of FESP on sequences with different lengths or on sequences with some characters missing, we randomly adopt several sequences in HS3D and cut off their last 1 to 6 characters. Experimental results of classification on the modified dataset are shown in Fig.6.

From Fig.6, we know that no matter the unequal lengths data set are produced by the training data or testing data, FESP can still implement the classification task and the accuracy of classifying the data set with different lengths or omitted characters is not obvious change.Feature selection for sequences with different lengths can be implemented by adding some common characters to build sequences with equal length.



	Raw data	training data with different lengths	testing data with different lengths	training and testing data with different lengths
■ min_S=0.03,k=300	0. 645004	0. 658002	0. 645004	0. 658814
■ min_S=0.05,k=60	0. 64094	0. 62795	0. 64257	0. 62713

Figure 6. Classification results for sequences with different lengths

V. CONCLUSION

Compared with Support Vector Machine (SVM) [10-11] and Neural Network, frequent emerging sequence pattern based classification algorithm shows the following advantages: (1)the algorithm keep the information provided by the order of bases in gene sequences for sequence classes and apply the information to classify new sequence; (2) the algorithm can catch interaction among bases and provide a clear description for causal relationship between DNA functions, so it is easy to understand; (3) when length of sequences in the DNA sequence database is not fixed or there exists some omitted base, the algorithm can be used directly without any modification; (4) some fragments of gene sequences carry more information of classes, the feature selection methods can find out these fragments, and experiments show that only using these fragments for classifying can achieve a higher accuracy.

ACKNOWLEDGMENT

This research was supported by the National Science Foundation in China No. 61070020 and the Program for New Century Excellent Talent in Fujian Province (XSJRC2007-11).

REFERENCES

[1] I.Guyon, J.Weston, S.Barnhill, V. Vapnik,2002. "Gene selection for cancer classification using support vector machines", *Machine Learning*. Volume 46, Issue 1-3:389-422.

[2] B. Liu, W. Hsu and Y. Ma. "Integrating classification and association rule mining", *Proceedings of ACM International Conference on Knowledge Discovery and Data Mining* (SIG KDD'98), pages 80-86, New York City, NY, August 1998.

[3] W. Li, J. Han and J. pei. "CMAR: Accurate and efficient classification based on multiple classification rules", *Proceedings of the 2001 IEEE International Conference on Data Mining*, California, 2001.

[4] G. Dong , X. Zhang , L. Wong and J. Li, "CAEP: Classification by Aggregating Emerging Patterns", *Proceedings of the Second International Conference on Discovery Scienc*e, p.30-42, December 01, 1999.

[5] O.R.Zaïane and M.L.Antonie, "Classifying text documents by associating terms with text categories", *Proceedings of thirteenth Australasian Database Conference* (ADC'02), pages 215-222, Melbourne, Australia, January 2002.

[6] G.Dong. and J.Li, "Efficient mining of emerging patterns: Discovering trends and differences", *Proceedings of International Conference on Knowledge Discovery and Data Mining* (KDD'99), pages43-52,San Diego, CA, Aug.1999.

[7] H.Liu and R.Setiono, "Chi2: Feature selection and discretization of numeric attributes", *Proceeding of IEEE 7th International Conference on Tools with Artificial Intelligence*, 338-391, 1995.

[8] Y. Xiong and Y. Zhu, "A multi-supports-based sequential pattern mining algorithm", In *Computer and Information Technology, 2005. CIT 2005. The Fifth International Conference on*, pages 170–174. IEEE, 2005.

[9] P.Pollastro and S.Rampone, "HS3D: Homo Sapiens Splice Site Data Set", Nucleic Acids Research, 2003 Annual Database Issue.

[10] V. Vapnik The Nature of Statistical Learning Theory. 2nd edition, NY: Springer.

[11] X. Zhang, X. Xiao and G. Xu, "Fuzzy Support Vector Machine Based on Affinity Among Samples", *Journal of Software*, 2006,17(5): 951-958.

Xiaoyun Chen received the BSc degree in applied mathematics and the Msc in computational mathematics from Fuzhou University of China in 1992 and 2001, respectively, and received PhD degrees in computer science from Fudan University of China in 2005. She is an associate professor of Information and Computational Science department at Fuzhou University of China. Her research interests include topics in text and time series and image data mining, pattern recognition.

Jinhua Chen received the BSc degree in applied mathematics from Fuzhou University of China in 2008. Currently, he is an Msc Candidate at the Fuzhou University of China. His main research interests are in anomaly detection and clustering algorithms.