Journal of the Korean
Data & Information Science Society
2006, Vol. 17, No. 4, pp. 1279~1287

Protein Motif Extraction via Feature Interval Selection[1)]

Insuk Sohn[2)] · Changha Hwang[3)] · Junsu Ko[4)]
David Chiu[5)] · Dug Hun Hong[6)]

Abstract

The purpose of this paper is to present a new algorithm for extracting the consensus pattern, or motif from sequence belonging to the same family. Two methods are considered for feature interval partitioning based on equal probability and equal-width interval partitioning. C2H2 zinc finger protein and epidermal growth factor protein sequences are used to demonstrate the effectiveness of the proposed algorithm for motif extraction. For two protein families, the equal width interval partitioning method performs better than the equal probability interval partitioning method.

Keywords : Feature interval selection, Protein motif, Support vector machine

1. Introduction

Mining of sequence data has many real world applications. Transaction history of a bank customer, product order history of a company, stock market and

1) This work was supported by the Korea Research Foundation Grant (KRF-2004-042-C00020).

2) Department of Statistics, Korea University, Seoul 136-701, Korea, E-mail : sis46@knu.ac.kr

3) Division of Information and Computer Science, Dankook University, Seoul 140-714, South Korea.

4) Object Interaction Technologies Inc, Daejeon, Korea

5) Department of Computing and Information Science, University of Guelph, Guelph, Canada N1G 2W1.

6) Corresponding Author : Department of Mathematics, Myongji University, Kyunggido 449-728, South Korea.
Email : dhhong@mju.ac.kr

biological DNA data are all sequence data where data mining techniques can be applied(Han and Kamber, 2001). In contrast to ordinary data, sequence data are dynamic and order dependent.

A protein sequence motif, signature or consensus pattern is a short sequence that is embedded within the sequences of a same protein family. Based on motif, an unknown sequence can be quickly classified into its computationally predicted protein family/families for further biological analysis. In past years, many algorithms for finding protein sequence motifs have been proposed. Some studies tackle the problem of protein motif identification using artificially generated data(Pevzner and Sze, 2000; Sagot, 1998; Buhler and Tompa, 2001) and some use real biological data to test against their algorithms(Chang and Halgamuge, 2002; Hart et al., 2000; Rigoutsos and Floratos, 1998; Smith and Smith, 1990).

The purpose of this paper is to extract relevant intervals using support vector machines(SVM) for extracting the consensus pattern or motif from sequence belonging to the same family. The paper is organized as follows. In section 2, we present the method of feature interval selection using SVM. In section 3, the capability of the proposed algorithm is demonstrated by using C2H2 zinc finger protein and epidermal growth factor protein sequences. Conclusions are given in section 4.

2. Method of Feature Interval Selection Using SVM

This section presents the methods of feature interval selection using SVM to extract relevant interval that best describes the dataset for the classifier generated. The proposed model of classification using feature interval selection is a wrapper model(Chiu and Leung, 1998; John et al., 1994), because the output result of the classifier is used as the measure for feature interval selection. Since the learning starts off with all features and removes feature one by one, the feature traversal is classified as sequential backward feature generation model(Koller and Sahami, 1996; Liu, 1998). From each feature, intervals are removed based on an appropriate ranking measure. Thus this phase can be a sequential backward model as well. The SVM classifier is used as the learning algorithm.

2.1 The Wrapper Model in Feature Interval Selection

Algorithms for feature selection problem usually fall into two categories, the filter(Hall, 2000) or the wrapper(John et al., 1994; Chiu and Leung, 1998) model. The filter model is independent of the learning algorithm and relies solely on the evaluation of the data without introducing the results from classification. On the other hand the wrapper model is dependent on the classification algorithm to make a decision whether a selection is appropriate or not. This implies that the

classification process is a contributing factor in the selection process. Here the selection of feature intervals is based on the output of the classification process. The advantage of this approach is that the behavior of the selection process is closely related to the classification process, and their performances of the two processes reinforce each other.

As in Messer and Kittler(1997), the feature intervals are selected based on the misclassification of the SVM classification. If the criterion for selection is satisfied by the misclassification of the SVM classification, then the feature set with the identified intervals learned by the SVM classifier will be selected. Otherwise it will be marked as irrelevant.

In advance, we briefly describe our method to be proposed as follows:

Feature Interval Selection Algorithm

1. Normalize value in each feature of the dataset.
2. Identify the intervals for each feature given a number of intervals to be considered.
3. Calculate the correlation coefficient between feature-pairs to rank the features to be considered.
4. Select a feature for interval evaluation.
5. Select an interval of the feature for evaluation.
6. Process data using the SVM classifier.
7. Compare the misclassification error: if the misclassification error decreases then the interval is considered for removal; otherwise the interval is retained.
8. Go back to 5 until all intervals within feature are processed.
9. Go back to 4 until all features are processed.

2.2 Normalization

The dataset is preprocessed in two ways. First the inputs are normalized and then an initial ranking is performed so that it can be used as a measure for feature selection. If each data sample has the form $X=\{f(1), f(2),\cdots, f(n)\}$, then the normalized value of each attribute $f(i)$ is equal to:

$$x(i)=\frac{f(i)-MIN(i)}{MAX(i)-MIN(i)}\times(MAXRANGE-MINRANGE)+MINRANGE$$

where $MAX(i)$ is the maximum value of feature i among all samples, and $MIN(i)$ is the minimum value of feature i. $MINRANGE$ is the minimum value of the interval that we want to scale it to. In the case of the SVM classifier the choice for MINRANGE and MAXRANGE is -1 and 1 respectively.

2.3 Feature Ranking

Initial ranking of features is important for the sequential traversal in the search process. Pearson's correlation matrix(Hall, 2000) is used to calculate the initial ranking of the features. A correlation matrix comparing each feature pair gives a statistical measure of the correlation among the features. Since correlation coefficients that are closer to 1 or −1 indicate a stronger statistical correlation of the feature pair, this particular feature pair should not be removed first(Hall, 2000). Feature values with a correlation value closer to 0 are better initial candidates for removal, since these features reflect less global characteristic, and thus less relevant for classification. To calculate Pearson's correlation matrix of two features are taken at a time and a correlation factor r for those two features is calculated:

$$r(x(i),\ x(j)) = \frac{1}{n}\sum\left(\frac{x_k(i)-\mu_i}{s_i}\right)\left(\frac{x_k(j)-\mu_j}{s_j}\right)$$

where μ's are the means and s's are the standard deviations of the features. The resulting factor r is in the range of −1 and 1, but often r^2 is used as a final correlation factor, because it does not differentiate between positive and negative association between the features.

2.4 Interval Selection

At an iteration, one interval of a feature is removed for the classification task, given that the feature values are partitioned into intervals. Two methods are considered for feature interval partitioning, referred to as equal probability(or maximum entropy) and equal-width interval partitioning(Wong and Chiu, 1987). The number of intervals to choose is dependent on the computational resources available such as the sample size. In results, but if the resources are not available then a number of intervals should be chosen consistent with the resources available. We choose the number of intervals for each feature as related to a multiple of the number of classes in the dataset assuming that the sample size is sufficient for the analysis. This means that if there are five different classes in the dataset and the multiple is set to one, then there is five times one interval. If the number of intervals was selected to be two then there will be ten different intervals(or two times five classes). Taking this approach, linearly separable classes that can be partitioned into intervals are more likely.

Equal-Probability Interval Partitioning

Dividing the data values with equal probability will produce intervals that contain the same number of data samples(Wong and Chiu, 1987). This means that

the intervals may not be of the same width and that they may vary from very small to very large, depending on the distribution of the dataset. If the data is uniformly distributed the intervals are the same as the equal width partitioning method. The intervals then contain the same number of sample points. The reason for using equal probability intervals is that the entropy for all intervals is being maximized. The entropy function after equal probability partitioning can be calculated below:

$$H_n(x(i)) = \sum_{k=1}^{n} p_k(x(i)) \log_2 \frac{1}{P_k(x(i))}$$

where $P_k(x_i)$ is the probability estimate for the interval k given that n is the number of predefined intervals and x_i is a feature to be discretized. Maximizing the feature entropy minimizes the information loss due to partitioning(Wong and Chiu, 1987), which is desirable when interval partitioning is needed. This is appropriate if class information should not be used in the partitioning to give a set of feature-dependent only intervals.

Equal Width Interval Partitioning

This method partitions the feature values into intervals by dividing the feature space into even intervals, independent of the sample distribution. If the data are distributed uniformly then the number of samples in each interval will be exactly the same. That is, both equal-width and equal-probability partitioning will produce the same intervals if the data are evenly distributed. This method of interval partitioning was chosen as a common approach to show the differences in performance of those two different methods. However, this method would split a feature into intervals with different number of samples, which means that those with more samples would contain more information than those with less samples

Ranking of Feature Interval for Selection

The ranking of the intervals of a feature for selection is calculated using the Shannon's entropy function on the class distribution within an interval. The higher is the value, the higher is the ranking for selection. That is, a small class entropy value is more useful for classification. This method is used so that the feature values are more likely to be associated with some of the classes. The class entropy function is defined below:

$$H_{x_i}(C) = \sum_{C} p_{x_i}(C) \log_2 \frac{1}{P_{x_i}(C)}$$

where $P_{x(i)}(C)$ is the probability estimate of the class within the interval X_i.

Substitution

After the removal of intervals, data values from the interval are substituted by values that the SVM classifier ignores in the learning process. This is accomplished by setting those interval values to 0. Even though this substituted value can be an observed zero from the dataset, the behavior of the SVM classifier is the same without an effect on the output.

3. Result

This section illustrates the performance of the proposed algorithm using two protein families of C2H2 zinc finger protein and EGF protein.

C2H2 Zinc Finger Protein

C2H2 are nucleic acid-binding protein structures first identified in the Xenopus transcription factor TFIIIA. These domains have been found in nucleic acid binding proteins. A zinc finger domain is composed of 25 to 30 amino-acid residues. There are two cysteine or histidine residues at both extremities of the domain, which are involved in the tetrahedral coordination of a zinc atom(PROSITE website). The C2H2 zinc finger protein has 418 sequences and the non-C2H2 finger protein has 453 sequences. The training dataset consists of the sequence of 279 C2H2 zinc finger protein and the sequence of 302 non-C2H2 finger protein and the test dataset consists of the sequence of 139 C2H2 zinc finger protein and the sequence of 151 non-C2H2 zinc finger protein.

C-x(2)-C-(12)-C-x(3)-H is selected the same preliminary motif as Chang and Halgamuge(2002). Each of the protein sequences is searched for the `most similar match' with the preliminary pattern and matching patterns from the sequences are used as the data for extracting the motif. The degree of similarity is obtained by calculating the sum of square error(SSE). More details of preliminary motif selection and SSE calculation can be found in Chang and Halgamuge(2002).

To analyze the results due to feature interval selection, different number of intervals is evaluated. The classification results are shown in Table1. Equal width interval partitioning method selected the motif pattern, C-x(2,4)-C-x(10,14)-H-x(2,5)-H, and correctly identified 128 out of 139 sequences(92%). However, it also incorrectly identified 8 non-C2H2 Zinc Finger protein sequences(5%). Equal probability interval partitioning method with 2 number of interval selected the motif pattern, C-x(1,4)-C-x(10,14)-H-x(1,5)-H, and correctly identified 99 out of 139 sequences (71%). However, it also incorrectly

identified 20 non-C2H2 Zinc Finger protein sequences(13%). For equal probability interval partitioning method, smaller number of intervals gives a better classification result. This result implies that the equal width interval partitioning method perform better than the equal probability interval partitioning method.

<Table 1> Classification performance of optimized motif patterns from C2H2 Zinc Finger Proteins. Max SSE is the maximum sum of square error.

Intervals per feature	Methods	Motif	True positives	False positives	Max SSE
2	Equal-width	C-x(2,4)-C-x(10,14)-H-x(2,5)-H	128/139 (92%)	8/151 (5%)	5
	Equal-prob	C-x(1,4)-C-x(10,14)-H-x(1,5)-H	99/139 (71%)	20/151 (13%)	
4	Equal-width	C-x(2,4)-C-x(10,14)-H-x(2,5)-H	128/139 (92%)	8/151 (5%)	
	Equal-prob	C-x(2,4)-C-x(11,14)-H-x(2,5)-H	87/139 (63%)	6/151 (4%)	

EGF Protein

The functional significance of epidermal growth factor(EGF) domains in what appear to be unrelated proteins is not yet clear. However, a common feature is that these repeats are found in the extracellular domain of membrane bound proteins or in proteins known to be secreted(exception: prostaglandin G/H synthase) (PROSITE website). The training dataset consist of the sequence of 297 EGF protein and the sequence of 2290 non-EGF protein and the test dataset consist of the sequence of 149 EGF protein and the sequence of 1645 non-EGF protein.

C-x(1)-C-x(5)-G-x(2)-C is selected the same preliminary motif as Chang and Halgamuge(2002). Each of the protein sequences is searched for the 'most similar match' with the preliminary pattern and matching patterns from the sequences are used as the data for extracting the motif. The degree of similarity is obtained by calculating the sum of square error(SSE).

To analyze the results due to feature interval selection, different number of intervals is evaluated. The classification results are shown in Table 2. Equal-width interval partitioning method with 2 number of interval selected the motif pattern, C-x(1,4)-C-x(2,8)-G-x(2,5)-C, and correctly identified 91 out of 149 sequences(61%). Equal probability interval partitioning method with 2 number of interval selected the motif pattern, C-x(1,4)-C-x(4,8)-G-x(1,5)-C, and correctly identified 24 out of 149 sequences (16%). For two interval partitioning method, smaller number of intervals gives a better classification result. This result implies that the equal width interval partitioning method perform better than the equal probability interval partitioning method.

<Table 2> Classification performance of optimized motif patterns from EGF Proteins. Max SSE is the maximum sum of square error allowed.

Intervals per feature	methods	Motif	True positives	False positives	Max SSE
2	Equal-width	C-x(1,4)-C-x(2,8)-G-x(2,5)-C	91/149 (61%)	0/1645 (0%)	10
	Equal-prob	C-x(1,4)-C-x(4,8)-G-x(1,5)-C	24/149 (16%)	0/1645 (0%)	
4	Equal-width	C-x(1,4)-C-x(2,8)-G-x(2,5)-C	91/149 (61%)	9/1645 (0.54%)	
	Equal-prob	C-x(1,4)-C-x(2,8)-G-x(2,5)-C	27/149 (18%)	8/1645 (0.48%)	

4. Conclusions

The purpose of this paper is to extract relevant intervals using support vector machines for extracting the consensus pattern or motif from sequence belonging to the same family. Two approaches for feature interval partitioning based on equal probability and equal width partitioning are chosen. C2H2 zinc finger protein and epidermal growth factor protein sequences are used to demonstrate the effectiveness of the proposed algorithm in motif extraction. For two protein families the equal width interval partitioning method performs better than the equal probability interval partitioning method and smaller number of intervals gives a better classification result.

References

1. Buhler, J. and Tompa, M. (2001). Finding motifs using random projections, *Proceedings of RECOMB 2001*, 69-76.
2. Chang, B. C. and Halgamuge, S. K. (2002). Protein motif extraction with neuro-fuzzy optimization, *Bioinformatics*, 18(8), 1084-1090.
3. Chiu, D. K. Y. and Leung M. Y. W. (1998). Evaluating input dependency in feedforward neural network, *5th International Conference on Neural Information Processing (ICONIP' 98)*, 801-804.
4. Hall, M. A. (2000). Correlation-based feature selection for discrete and numeric class machine learning, *Proceedings of the 17th International conference.*
5. Han, J. and Kamber, M. (2001). *Data Mining: Concepts and Technigues*, Academic Press.
6. Hart, R., Royyuru, A., Stolovitzky, S. and Califano, A. (2000). Systematic and automated discovery of patterns in PROSITE families, RE*COMB 2000*, 147-154.

7. John, G. H., Kohavi, R. and Pfleger (1994). Irrelevant features and the subset selection problem, *Proceedings of the 11th International Conference on Machine Learning ICML94*, 121-129.
8. Koller, D. and Sahami, M. (1996). Toward optimal feature selection, *Proceedings of the 13th International conference*, 284-292.
9. Liu, H. (1998). *Feature Selection for Knowledge Discovery and Data Mining*. Kluwer Academic Publishers.
10. Messer, K. and Kittler, J. (1997). A comparison of colour texture attributes selected by statistical feature selection and neural network methods, *Pattern Recognition Letters*, 1241-1246.
11. Pevzner, P. and Sze, S. (2000). Combinatiorial approaches to finding subtle signals in DNA sequences, *Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology*, 269-287.
12. PROSITE database website, http://www.expasy.ch/sprot/sprot/sprot-top.html.
13. Rigoutsos, I. and Floratos, A. (1998). Motif discovery without alignment or enumeration, *RECOMB 98*, 221-227.
14. Sagot, M. (1998). Spelling approximate repeated or common motifs using a suffi tree, *Lecture Notes in Computer Science*, 1380, 111-127.
15. Smith, R. and Smith, T. (1990). Automatic generation of primary sequence patterns from sets of related protein sequences, *Nucleic Acids Res.*, 118-122.
16. Wong, A. K. C. and Chiu, D. K. Y. (1987). Synthesizing statistical knowledge from incomplete mixed-mode data, *IEEE trans. on Pattern Analysis and Machine Intelligence*, 9(6), 796-805.

[received date : July. 2006, accepted date : Sep. 2006]