HIERARCHICAL FUZZY CLASSIFIER FOR BIOINFORMATICS DATA[▽]

CHONG, A.[1], GEDEON, T.D.[1], KOCZY, L.T.[1,2]

[1]School of Information Technology
Murdoch University
Western Australia
Email: {cchong,gedeon}@murdoch.edu.au

[2] Department of Telecom & Telematics
Budapest University of Technology and Economics
and
Institute of IT and Electrical Engineering, Széchenyi István University, Győr
HUNGARY
Email: koczy@ttt.bme.hu

ABSTRACT

In this research, a preliminary study of the application of hierarchical fuzzy rule-based classifier for protein secondary structure prediction has been carried out. The use of a hierarchical structured rulebase alleviates, to some extent, the problem of rule explosions that has prevented the use of traditional fuzzy system in many biomedical related problems. As part of the study, a hierarchical fuzzy classifier was built from a set of training data. Although the accuracy of the classifier is far from comparable to the current established techniques, the experiment has successfully confirmed the feasibility of the application of the hierarchical classifier for protein structure prediction. This calls for further research to further improve the accuracy of the rule-based classifier. The advantages of using the rule-based classifier as compared to other artificial intelligent techniques for protein structure prediction are also discussed in the paper.

1 INTRODUCTION

Fuzzy rule-based classifier has attracted a significant amount of attention from researchers in the field of pattern recognition and digital signal processing. They distinguished themselves from other artificial intelligence techniques by not only producing reliable conclusions, but also allowing users to interpret how the conclusions are inferred. Unfortunately, the advantage comes with a price. In general, the number of rules needed to cover the entire problem domain is $O(T^K)$ where K is the number of features (dimensions) and T is the number of terms per input. In other words, the numbers of rules grows exponentially as the number of input features increases.

In the field of biomedical, or more specifically bioinformatics, artificial intelligence-based classifiers have been applied for several problems. One of such problems is the problem of protein structure prediction. The structure of a particular protein determines its function. The protein structure information is, therefore, useful in a wide range of biomedical related fields. Protein structure determination by physical experimentation is a time and resource-consuming task. Over the years, many new proteins have been identified by large-scale genome sequencing projects. While the protein sequence of the new protein can be identified, the protein structure is often not known. As an attempt to narrow the gap between the number of known protein sequences and the number of experimentally determined protein structures, methods for protein structure prediction have been studied [1, 2].

Among the artificial intelligent techniques explored for protein structure prediction, neural networks have been predominant in the literature. The main drawback of neural networks is that they operate in a black box manner. In this aspect, fuzzy rule-based classifier is potentially more superior to neural network. A review of literature suggests that there is almost no report on the application fuzzy rule-based classifier for protein structure prediction. This is mainly due to the large number of features (system inputs) necessary to accomplish the task. A neural network designed for this purpose usually use up to 130 – 260 inputs. It is hardly possible to construct a traditional fuzzy rule-base classifier that operates on such large number of inputs.

A hierarchical fuzzy classifier can potentially alleviate the problem of rules explosion in fuzzy rule-based classifier [3]. In this research, we attempt to apply the hierarchical fuzzy classifier for protein secondary structure prediction. The aim is not to generate a classifier that outperforms well-established techniques reported in the past, but to serve as a preliminary investigation on the application of fuzzy rule-based classifier in protein secondary structure prediction.

▽ Research supported by the the Australian Research Council, National Scientific Research Fund OTKA T034233 and T034212, a Main Research Direction Grant 2002 by Széchenyi István University, and the National Research and Development Project Grant NKFP-2/0015/2002.

2 PROTEIN SECONDARY STRUCTURE PREDICTION

The protein structure prediction is normally performed by observing the protein sequence combined with our prior knowledge on a set of homologous proteins whose structure has been determined. At the time of writing, the prediction of a protein's three-dimensional structure from its amino acid chain (protein sequence) remains an unsolved problem. A review of the literature suggests that most of the research in this problem domain addresses the prediction of protein secondary structure. Most of the work done attempts to predict a protein sequence to be one of the following: Helix (H), Extended (E) and Loop (L). The problem can be viewed as a simple classification problem. Given a protein sequence, some algorithms are applied to classify the protein structure into Helix, Extended or Loop.

The input to the protein structure classifier is usually a 'window' of residues from the chain of amino acid that composes the protein. Given a protein sequence, a window of n (usually 7 or 13) residues is examined at a time and the secondary structure of the amino acids in the middle (position 7 for the window size of 13) is predicted.

It has been reported in [4] that the use of profile information resulted from sequence alignment can improve the classification accuracy. For each input protein sequence, the sequence alignment process involves the use of bioinformatics tools (e.g. BLAST) to search for homologous protein sequences in the SWISSPROT data bank. The output of the program is a list of homologous protein sequences as well as their similarity (in percentages) to the input sequence.

To include the profile information in the classification, each residue in the window is represented by 20 entries of profile information. From the data representation point of view, a window size of 13 results in 13 x 20 = 260 input features. In this case, the complexity of a traditional fuzzy rule-based classifier will be $O(T^{260})$!

3 FUZZY CLUSTERING

Given a set of data, Fuzzy c-Means clustering (FCMC) [5] performs clustering by iteratively searching for a set of fuzzy partitions and the associated cluster centers that represent the structure of the data as best as possible. Given the number of clusters c, FCMC partitions the data $X = \{x_1, x_2, \ldots, x_n\}$ into c fuzzy partitions by minimizing the within group sum of squared error objective function as follows (Eqn 1).

$$J_m(U,V) = \sum_{k=1}^{n}\sum_{i=1}^{c}(U_{ik})^m \| x_k - v_i \|^2,\ 1 \le m \le \infty \qquad \text{Eqn 1}$$

where $J_m(U,V)$ is the sum of squared error for the set of fuzzy clusters represented by the membership matrix U, and the associated set of cluster centers V. Here, $\|x_k - v_i\|^2$ represents the distance between the data x_k and the cluster center v_i. At each iteration, the cluster centers are calculated using (Eqn 2) and (Eqn 3).

$$U_{ik} = \left(\sum_{j=1}^{c} \left(\frac{\| x_k - v_i \|}{\| x_k - v_j \|} \right)^{2/(m-1)} \right)^{-1} \quad \forall i, \forall k, \qquad \text{Eqn 2}$$

and

$$v_i = \frac{\sum_{k=1}^{n} (U_{ik})^m x_k}{\sum_{k=1}^{n} (U_{ik})^m}, \qquad \text{Eqn 3}$$

The optimal number of clusters in the data is determined by means of the FS index [6] as follows:

$$S(c) = \sum_{k=1}^{n}\sum_{i=1}^{c}(U_{ik})^m(\|x_k - v_i\|^2 - \|v_i - \bar{x}\|^2) \quad 2<c<n, \qquad \text{Eqn 4}$$

The number of cluster, c, is determined so that $S(c)$ reaches a local minimum as c increases.

4 FUZZY RULEBASE CLASSIFIERS

The fuzzy rulebase is one of the most successful applications of fuzzy logic. It has been used in a wide range of application domain. Depending on the problem domain, fuzzy rulebases are called different names, such as fuzzy classifiers (for classification), fuzzy controllers (robotics and control), and more generally fuzzy systems. In this paper, the terms are used interchangeably. Fuzzy systems operate on fuzzy rules of the form:

If X is A_i then Y is B_i

Here, $X = \{x_1, x_2, \ldots, x_n\}$ is the input, Y is the output, $A_i = A_{i1} \times A_{i2} \times \ldots \times A_{in}$ and B_i are fuzzy sets of the antecedent and the consequent of the rule respectively. Each rule defines a "patch", (i.e. a fuzzy relation) in the $(X \times Y)$ space. The number of rules (patches) necessary to cover the entire problem domain grows exponentially with the increased number of features. To alleviate this problem, a hierarchical fuzzy system is proposed [3].

The hierarchical fuzzy system is based on the following idea. Often, the multi-dimensional input space $X = X_1 \times X_2 \times \ldots \times X_k$ can be decomposed into some subspaces, e.g. $Z_0 = X_1 \times X_2 \times \ldots \times X_{k0}$ $(k_0 < k)$, so that in Z_0 a partition $\Pi = \{D_1, D_2, D_3\}$ can be determined. In each D_i, a sub-rule base R_i can be constructed with local validity. The hierarchical rule base structure becomes:

R_0: if Z_0 is D_1 then use R_1
if Z_0 is D_2 then use R_2 ... ...
... ...
if Z_0 is D_n then use R_n R_n: if Z_n is A_{n1} then y is B_{n1}
... ...
R_1: if Z_1 is A_{11} then y is B_{11} if Z_n is A_{nmn} then y is B_{nmn}
... ...
if Z_1 is A_{1m1} then y is B_{1m1}

The complexity of fuzzy systems can be reduced when the suitable Z_0 and Π are found such that in each sub-rule base R_i the input space X_i is a subspace of $X / Z_0 = X_{k0+1} \times X_{k0+2} \times \ldots \times X_k$ [3].

5 HIERARCHICAL FUZZY

In this research, we attempt to automatically construct a hierarchical fuzzy classifier from the set of data obtained from [4]. The window size of 7 is chosen (see section 2 on details about window size). This results in a set of 140 dimensional data. Altogether 2400 data points are used, from which 2000 are used for training and the remaining 400 for testing.

The algorithm in [7] is used for the automatic construction of the hierarchical classifier. Firstly, each dimension is ranked by its suitability in forming Z_0 for meta-rules construction. This is done by projecting the data points to each dimension and performing fuzzy clustering [5] (section 3) to obtain one-dimensional clusters. The importance (suitability) of each dimension is ranked by observing the partition matrix $U = \{\mu_{ik} \mid i = 1 \ldots N_c,\ k = 1 \ldots N\}$ resulted from clustering and using the interclass separability criterion:

$$J(X') = \frac{tr(Q_b)}{tr(Q_w)} \qquad \textbf{Eqn 5}$$

$$Q_b = \sum_{i=1}^{N_c} \sum_{j=1}^{N} \mu_{ij}^m (v_i - \bar{v})(v_i - \bar{v})^T \qquad \textbf{Eqn 6}$$

$$v_i = \frac{\sum_{k=1}^{n} (U_{ik})^m x_k}{\sum_{k=1}^{n} (U_{ik})^m} \qquad \textbf{Eqn 7}$$

$$Q_i = \frac{1}{\sum_{j=1}^{N} \mu_{ij}^m} \sum_{j=1}^{N} \mu_{ij}^m (x_j - v_i)(x_j - v_i)^T \qquad \textbf{Eqn 8}$$

$$Q_w = \sum_{i=1}^{N_c} Q_i \qquad \textbf{Eqn 9}$$

$$\bar{v} = \frac{1}{N_c} \sum_{i=1}^{N_c} v_i \qquad \textbf{Eqn 10}$$

In (Eqn 5), '*tr*' denotes the trace of a matrix. The goal is to choose a subset of inputs whose value in (Eqn 5) is relatively larger than the others. Figure 1 shows the 140 input features sorted by separability (descending order) and their corresponding separability value. In [7], the number of features for Z_0 construction is chosen by creating multiple hierarchical fuzzy systems and selecting the one with the best performance. The proposed approach is less feasible in this experiment due to high computational time required. In this experiment, the number of features is determined by subjective observation from Figure 1. The first three top-ranked variable is selected due to the relatively bigger jump in the separability value for the fourth variable. The actual subspace obtained is $Z_0 = X_{28} \times X_{108} \times X_{128}$.

Next, fuzzy clustering is performed on the data along subspace Z_0 and a fuzzy partition with ten clusters $\Pi = \{D_1, \ldots, D_{10}\}$ were obtained. From the fuzzy partition Π, a crisp partition of the data points is constructed. That is, for each fuzzy cluster D_i, the corresponding crisp cluster of points is determined as $P_i = \{p \mid \mu_i(p) > \mu_j(p)\ \forall\ j \neq i\}$. The crisp clusters of points then become the target for automatic fuzzy rule extraction. A sub-rulebase R_i can be constructed using any rule extraction techniques in the literature on P_i. In this research, the technique in [8] is used. In general, fuzzy clustering is performed on the data being modelled. Each cluster is then converted to one or more fuzzy rules. Upon obtaining R_i, a meta rule can be formed as: *If* Z_0 *is* D_i *then use* R_i

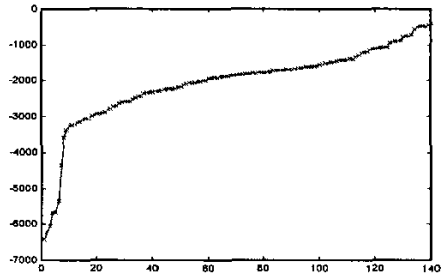

Figure 1. Input Features and their corresponding separability.

Sub-Rulebase	Input features used	# Rules
1	X_{119}, X_{127}	4
2	X_{43}, X_{74}	7
3	X_{11}, X_{67}	4
4	$X_{13}, X_{17}, X_{22}, X_{26}, X_{34}, X_{92}, X_{131}$	10
5	X_{11}	4
6	X_1, X_{64}	3
7	X_{131}	7
8	X_{63}	5
9	$X_{120}, X_{109}, X_{117}, X_{103}$	4
10	X_{91}	3

Table 1. Sub-rulebase number of rules and input features

Prior to the rule extraction, the data points go through a feature selection process to eliminate unimportant input features [9]. The process involved fuzzy clustering the data points into n clusters and selecting the features that maximizes (Eqn 5). The remaining input features are then used in the rule extraction process. Table 1 shows the 10 sub-rulebases constructed and their corresponding number of rules as well as the input features used.

6 RESULTS AND DISCUSSION

In this section, the performance of the generated hierarchical fuzzy classifier is evaluated and discussed. The accuracy of the resulting model is calculated as the

percentage of correctly predicted residues. Since this is a preliminary investigation on the application of hierarchical fuzzy for structure prediction, it is not reasonable to expect accuracy comparable to well-established techniques in the literature. In the field of protein structure prediction, there is much discussion on the so-called 'trivial' system. A trivial system is a system that predicts only LOOP regardless of the system input. The main idea is that LOOP is the most often observed secondary structure, constituting sometimes up to 50% of the population on some datasets. At this stage, we aim at comparing our hierarchical model with only the 'trivial' system. Table 2 shows the accuracy of each individual sub-rulebase and the corresponding number of data points being modelled.

Sub-Rulebase	Accuracy	# Data Points
1	52.99%	117
2	55.06%	89
3	53.09%	194
4	47.52%	1248
5	54.55%	77
6	53.57%	84
7	50.00%	132
8	40.54%	74
9	54.55%	77
10	52.44%	82

Table 2. Sub-rulebase number of rules and input features

The overall accuracy of the hierarchical classifier on the training data is 47.61%. The accuracy of the trivial system is 46.55%. For the test set, the accuracy of the hierarchical classifier and the trivial system are 46.22% and 45.75% respectively.

It can be observed that the hierarchical fuzzy classifier was able to outperform the trivial system slightly in this experiment. It is, however, disappointing that the overall accuracy of the classifier is still very low (<50%). This shows that the hierarchical fuzzy classifier is far from comparable to established techniques, such as PHD [4] whose accuracy is greater than 70%. In this case, the accuracy of the fuzzy classifier is traded off for model interpretability. Much effort is needed to improve the hierarchical fuzzy classifier accuracy.

In general, the use of feature selection has reduced the number of features from 130 to 21, namely X_1, X_{11}, X_{13}, X_{17}, X_{22}, X_{26}, X_{34}, X_{43}, X_{63}, X_{64}, X_{67}, X_{74}, X_{91}, X_{92}, X_{103}, X_{109}, X_{117}, X_{119}, X_{120}, X_{127}, and X131. This is one of the main reasons for the low accuracy of the resulting classifier compared to neural networks that use all the features for classification. However, even with the reduced number of features, traditional fuzzy systems are hardly able to perform the task effectively. Most of the successful fuzzy systems are limited to handling only 5-10 features. The use of a hierarchical model has further reduced the system complexity by separating the problem domain into multiple sub-domains such that in each sub-domain, only a subset of the remaining variables plays a significant role in influencing the output. The final system complexity is $O(T^{10})$.

7 CONCLUSION

A preliminary investigation on the application of hierarchical fuzzy rule-based classifier for protein structure prediction has been carried out. As part of the study, a hierarchical fuzzy classifier was built from a set of training data. Although the accuracy of the classifier is far from comparable to the current established techniques, the experiment has successfully confirmed the feasibility of applying the hierarchical classifier for protein structure prediction. Further research is necessary to improve the accuracy of the rule-based classifier. Once the accuracy is improved, the rule-based classifier can be more superior than neural-network based classifiers due to its ability to explain how the conclusions are inferred.

References:

[1] T. R. Defay and F. E. Cohen, "Multiple sequence information for threading algorithms," *J. Mol. Biol.*, vol. 262, pp. 314-323, 1996.

[2] D. Fischer and D. Eisenberg, "Protein fold recognition using sequence-derived predictions," *Protein Sci.*, vol. 5, pp. 947-955, 1996.

[3] L. T. Koczy and K. Hirota, "Approximative inference in hierarchical structured rule bases," presented at Fift IFSA World Congress, Seoul, 1993.

[4] B. Rost and C. Sander, "Prediction of Protein Secondary Structure at Better than 70% Accuracy," *J. Mol. Biol.*, vol. 232, pp. 584-599, 1993.

[5] J. C. Bezdek, *Pattern Reconition with Fuzzy Objective Function Algorithms*. New York: Plenum Press, 1981.

[6] Y. Fukuyama and M. Sugeno, "A new method of choosing the number of clusters for fuzzy c-means method," presented at Proceedings of the 5th Fuzzy Systems Symposium, 1989.

[7] A. Chong, T. D. Gedeon, and L. T. Koczy, "Hierarchical Fuzzy Modelling," presented at International Fuzzy Association World Congress IFSA'03 (accepted), Istanbul, 2003.

[8] A. Chong, T. D. Gedeon, and L. T. Koczy, "Projection Based Method for Sparse Fuzzy System Generation," presented at 2nd WSEAS Int. Conf. on Scientific Computation and Soft Computing, Crete, 2002.

[9] D. Tikk and T. D. Gedeon, "Feature ranking based on interclass separability for fuzzy control aplication," presented at Proceedings of the International Conference on Artificial Intelligence in Science and Technology (AISAT'2000), Horbat, 2000.