A Fuzzy Rule-Based Classification System for classification of protein structural classes

Farzaneh M. Parizi1, Eghbal G. Mansoori
School of Electrical and Computer Engineering, Shiraz University, Shiraz, Iran
meimandi@cse.shirazu.ac.ir
Paper Reference Number: 07-04-0416

Abstract

For pattern classification in various fields, Fuzzy Classification Systems have been used, since they can improve performance of the classification system and generate an interpretable classification fuzzy system. Actually, a fuzzy system which generates a rule base with fewer and shorter general rules provides more comprehensible system. We have employed a FRBCS for classification of protein structural classes, since there is a need to implement reliable, accurate and comprehensible classification systems in this field. In this study we tested our method with a benchmark dataset and achieved good classification accuracy. The generated rule base is compact and comprehensible.

Key words: Fuzzy rule-based classification system, , protein structural classes, interpretability

1. Introduction

2. Fuzzy classification systems are a set of fuzzy "if-then" rules with linguistic variables, which are likely to be easily understood by human beings. They are often applied to classification problems in various fields. We have utilized a FRBCS for Classification of protein structural classes, since it is a challenging task in bioinformatics.
3. Protein structural classes' definitions were developed in 1976 and redefined multiple times thereafter. Researchers claim that the SCOP classification [1] is more natural and provides more reliable information to study protein structural classes. The SCOP classification currently includes eleven classes but he first four categories: (1) all-α proteins; (2) all-β proteins; (3) α/β proteins; (4) α+β proteins, include the great

[1] meimandi@cse.shirazu.ac.ir

majority of protein sequences and were the basis for most approaches, thus we focus on these four classes. Classification of protein structural classes has generally two aspects: feature vector and classification algorithm.

4. Existing features are extensively reviewed in [2]. Meanwhile several classification methods such as principal component analysis [3], Bayesian classifier [4], neural network [5], rough sets [6] and support vector machines [7], fuzzy k-nearest neighbor [8], fuzzy clustering [17, 18] and many other methods have been proposed in the literature.
5. Although several classification methods have been proposed, interpretability of the classification process hasn't received much attention in classification of protein structural classes. Involving fuzzy systems evolves this characteristic. In this study we employed a fuzzy classification system for prediction. Also we represented protein sequence as a feature vector with the information obtained from protein sequence and predicted secondary structure of protein. The resulted classification system achieved high prediction accuracy with jackknife cross-validation test for two benchmark datasets known as 640 dataset and 204 dataset.

6. Data and Material

I. Data

In this study we used two standard benchmark datasets 204 dataset and 640 dataset with 204 and 640 protein domains respectively, in order to compare our result with existing methods for classification of protein structural classes. The 204 dataset [9] has pair wise sequence identity lower than 40%, and includes 52 All-α class proteins, 61 All-β class proteins, 45 α/β class proteins and 46 α+β class proteins. The second one 640 dataset [10] with 25% sequence similarity, consists of 138 All-α class proteins, 154 All-β class proteins, 177 α/β class proteins and 171 α+β class proteins

II. Features

To classify protein structures, proteins are represented by numeric values known as features. We considered using features which are more familiar to biologists. Different kinds of protein representations are proposed by researchers. We found that most features which are extracted from secondary structure of proteins are more human understandable. Hence we predicted the secondary structure of protein sequences by PSIPRED [11] afterward used it to extract 19 features which were proposed in [12, 13, 14].

7. Research Methodology

The design of fuzzy rule-based classification systems consist of finding a compact set of fuzzy "if-then" classification rules to be able to model the input-output behavior of the system. Each Fuzzy rule is of the following form for n-dimensional pattern classification:
Rule R_j: If x_1 is A_{j1} and . . . and x_n is A_{jn}, then class C_j , for $j = 1, \ldots, N$.
According to Ishibuchi & Yamamoto [15] the comprehensibility of fuzzy rule-based systems is related to various factors :

a) Comprehensibility of fuzzy partitions (e.g., linguistic interpretability of each fuzzy set, separation of neighboring fuzzy sets, the number of fuzzy sets for each variable).
b) Simplicity of fuzzy rule-based systems (e.g., the number of input variables, the number of fuzzy if-then rules).
c) Simplicity of fuzzy if-then rules (e.g., type of fuzzy if-then rules, the number of antecedent conditions in each fuzzy if-then rule).
d) Simplicity of fuzzy reasoning (e.g., selection of a single winner rule, voting by multiple rules).

So that for having comprehensible fuzzy partitions the method used in this study partitions the problem space with usual K fuzzy sets on each dimension and with triangular membership functions. These membership functions for four different values of K are shown in Figure1.
If the paper is the result of a research, then the data and material used in the research should be presented here.

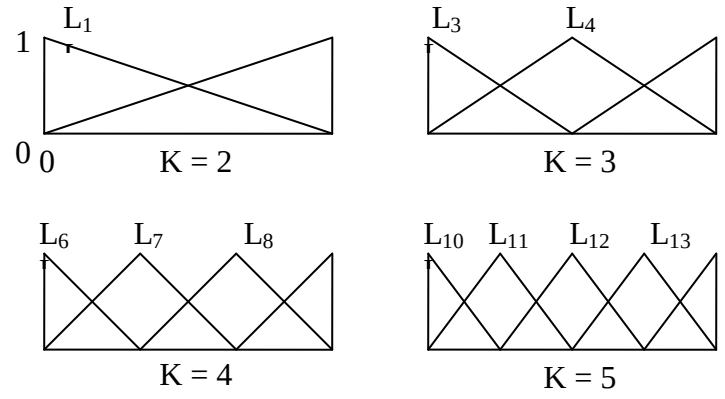

Figure 1: Different partitioning of each input attribute.

Moreover in order to achieve a simple fuzzy rule-based system, m fuzzy rules would be generated at most having a dataset with m data patterns. The third factor means the suitable number of antecedent conditions in each fuzzy if-then rule is specified through a generalization process. Besides to fulfill the fourth factor simple and intuitive method of fuzzy reasoning, single winner is used. The method generates only the rules that have at least one pattern in their decision subspace
In generalization phase, the antecedent's conditions of the rules are removed or replaced by more general sets. Generalizing a fuzzy set in the antecedent part of a rule will increase its covering area. To generalize a part of a rule in level K (where K fuzzy sets is defined on an attribute) its corresponding fuzzy set in level K-1 should be found. Suppose we have rule "Rj: If x1 is A1 and … xi is Ai and… xn is An then …" and Ai is a fuzzy set on level K and Ai' is its corresponding fuzzy set in level k-1. If we replace Ai with Ai' in Rj the new rule Rj' is more general than Rj. Eventually the method tries to find the best combination of the rules

with a hill climbing method, after generating and generalizing the rules. Accordingly, some of them are selected to form the final rule base.

4. Results and Analysis

In this section we evaluated the performance of the proposed method and feature vector with other computational methods for prediction of protein structural classes. Among cross validation methods, Jackknife cross validation test is usually used, since it yields a unique result for a given data set. Hence, all of the reported accuracies in this paper are obtained with jackknife cross validation test .The results of the proposed fuzzy classification system is shown in Table 1 for the two datasets. Also we reported the average number of rules generated in the rule bases of n-fold and also the average length of each rules generated in each rule base of each fold. The results show that the average length of rules are about 4 which show that rules are simple and the number of all rules are quite small for 204 dataset but larger for 640 dataset.

To fairly evaluate performance of proposed feature vector with other existing feature representations, we used the same classifier known as C4.5 decision tree which was used in [16]. The comparison of our proposed feature vector with the feature vector used in [16] using the same classifier is reported in table2. It is evident for two datasets that prediction accuracy of proposed feature vector is 12.7488 % higher on 204 dataset and 29.5288% higher on 640 dataset.

The existing fuzzy-based or rule-based classification methods for classification of protein structural classes are C4.5 [16], Fuzzy K-Nearest Neighbor [8] Unsupervised fuzzy clustering [17], Supervised fuzzy clustering [18], and Fuzzy SVM [19]. Comparison of our proposed method with other reported rule-based and fuzzy-based methods are shown in Table 3.

In order to have comprehensible pattern classification systems as well as having high prediction accuracy, fuzzy classification systems are being used. We proposed a fuzzy rule-based classification system which generated comprehensible rules. It seems, comprehensibility of classification systems hasn't been taken into account for classification of protein structural classes, although different classification systems have been proposed in the literature with good prediction accuracies. To do so we utilized human understandable features, most of them were extracted from predicted secondary structure of proteins.

Data set	All-α	All-β	α/β	α+β	Total	No. of rules	Avg. lentgh of rules
204	96.1538	100	100	97.826	98.5294	7.6373	4.0539
640	86.2319	80.5195	93.2203	51.462	77.5	28.4516	4.9279

Table 1. Performance of our method for 204 dataset and 640 dataset

Data set	Features	Accuracy(%)				
		All-α	All-β	α/β	α+β	Total
204	[16]	78.85	96.72	82.22	76.09	84.31
	This paper	98.1	95.1	100	95.7	97.0588
640	[16]	59.42	49.35	58.19	28.65	48.44

	This paper	89.9	82.5	78	64.3	77.9688

Table 2. Jackknife accuracy of our method (SGERD) with C4.5 on the same feature set

Data set	Method	Features	Accuracy (%)				
			All-α	All-β	α/β	α+β	Total
204	FuzzyKNN	Pseudo-amino acid composition	96.2	98.4	93.5	100	97
	Unsupervised fuzzy clustering	amino acid composition	67.3	86.9	60.9	46.7	68.1
	Supervised fuzzy clustering	amino acid composition	73.1	90.2	63.1	62.2	73.5
	Fuzzy SVM	Multi Pseudo-amino acid composition	92.3	100	82.6	93.3	92.6
	C4.5	PSI-BLAST based P-collocated amino acid pairs	78.85	96.72	82.22	76.09	84.31
	C4.5	This paper	98.1	95.1	100	95.7	97.06
	This paper	This paper	96.16	100	100	97.83	98.53
640	C4.5	PSI-BLAST based P-collocated amino acid pairs	59.42	49.35	58.19	28.65	48.44
	C4.5	This paper	89.9	82.5	78	64.3	77.97
	This paper	This Paper	86.23	80.52	93.22	51.46	77.5

Table 3 Comparison of Jackknife accuracies of the proposed method with other methods and feature sets

Thereafter we fairly examined the feature vector used by our method and the feature vector used in [16] with same C4.5 classifier and achieved higher prediction accuracy for the two datasets. This clarifies the feature set used by our method is better source of information.
Actually the proposed method generates compact rule base, the average number of rules and average length of rules generated are 7.6373 and 4.0539 respectively for 204 dataset and 28.4516 and 4.9279 respectively for the 640 dataset. In addition we fairly compared our method with existing fuzzy-based or rule-based classifiers for classification of protein structural classes and achieved higher prediction accuracy for the two datasets. For the 204 dataset FuzzyKNN [8] was the best performing in the mentioned existing methods with 97% overall prediction accuracy; however we achieved 98.5294% accuracy which is 1.5294% higher. Similarly for the 640 dataset that was used in [16] with C4.5 classifier, which has reported 48.44% but we achieved 77.5% prediction accuracy which is 29.06% higher.
However support vector machine classifier proposed in [12] has gained better prediction accuracies for 640 dataset in this field, still our proposed method has promising prediction accuracy as well as being human understandable. Naturally there is a tradeoff between interpretability and prediction accuracy.

5. Conclusions

In this work, we proposed a classification method and feature set that are comprehensible. The results obtained show that the fuzzy rule-based classification system used in this study is a method which grants interpretability to the prediction method also has a good performance for prediction of protein structural classes, and produces a compact rule base.

References

Murzin A.G., Brenner S.E., Hubbard T., Chothia C.(1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol*., 247, 536–540.

Kurgan L.A., Homaeian L.(2006). Prediction of structural classes for protein sequences and domains—impact of prediction algorithms, sequence representation and homology, and test procedures on accuracy. *Pattern Recognit.,* 39, 2323–2343.

Du Q.S., Jiang Z.Q., He W.Z., Li D.P., Chou K.C.(2006). Amino acid principal component analysis (AAPCA) and its applications in protein structural class prediction. *J. Bimol. Struct. Dynam*., 23, 635–640.

Wang Z.X., Yuan Z.(2000). How good is prediction of protein structural class by the component-coupled method? *Proteins*, 38 , 165-175.

Cai Y.D., Li Y.X., Chou K.C.(2000). Using neural networks for prediction of domain structural classes. *Biochim. Biophys. Acta*., 1476, 1–2.

Cao Y.F., Liu S., Zhang L., Qin J., Wang J., Tang K.X.(2006). Prediction of protein structural class with rough sets. *BMC Bioinformatics*, 7, 1–6.

Chao C., Xibin Z., Yuanxin T., Xiaoyong Z., Peixiang C.(2006). Predicting protein structural class with pseudo-amino acid composition and support vector machine fusion network. *Anal. Biochem*., 357, 116–121.

Zhang T.L., Ding Y.S., Chou K.C.(2008). Prediction protein structural classes with pseudo-amino acid composition: Approximate entropy and hydrophobicity pattern. *J. Theor. Biol.,* 250, 186-193.

Chou K.C.(1999). A key driving force in determination of protein structural classes. *Biochem. Biophys. Res. Commun*., 264, 216-224.

Chen K., Kurgan L.A., Ruan J.(2008). Prediction of protein structural class using novel evolutionary collocation-based sequence representation. *J. Comput. Chem*., 29, 1596–1604.

Jones D.T.(1999). Protein secondary structure prediction based on position-specific scoring matrices, *J. Mol. Biol*., 292, 195-202.

Kurgan L., Cios K., Chen K.(2008). SCPRED: accurate prediction of protein structural class for sequences of twilight-zone similarity with predicting sequences. *BMC Bioinform*., 9, 226.

Liu T., Jia C.(2010). A high-accuracy protein structural class prediction algorithm using predicted secondary structural information, *J. Theor. Biol*., 267, 272-275.

Zhang S.L., Ding S.Y., Wang T.M.(2011). High-accuracy prediction of protein structural class for low-similarity sequences based on predicted secondary structure. *Biochimie*, 93, 710–714.

Ishibuchi H., Yamamoto T.(2004). Fuzzy rule selection by multi-objective genetic local search algorithms and rule evaluation measures in data mining. *Fuzzy Sets Syst*., 141, 59–88.

Chen K., Kurgan L.A., Ruan J.(2008). Prediction of protein structural class using novel evolutionary collocation-based sequence representation. *J. Comput. Chem*., 29, 1596–1604.

Shen H.B., Yang J., Liu X.J., Chou K.C.(2005). Using supervised fuzzy clustering to predict protein structural classes. *Biochem. Biophys. Res. Commun*., 334, 577–581.

Zhang C.T., Chou K.C., Maggiora G.M.(1995). Predicting protein structural classes from amino acid composition: application of fuzzy clustering. *Protein Eng*., 8, 425–435.

Ding, Y.S., Zhang, T.L., Chou, K.C., 2007. Prediction of protein structure classes with pseudo amino acid composition and fuzzy support vector machines network. *Protein Peptide Lett.* 14, 811–815.