Feature Selection and Combination Methods for Protein Fold Classification*

Chiou-Yi Hor, Shyue-Horng Shiau and Chang-Biau Yang†
Department of Computer Science and Engineering
National Sun Yat-sen University, Kaohsiung, Taiwan
E-mail:cbyang@cse.nsysu.edu.tw

***Abstract*—In this paper, we propose a feature selection and combination method for protein fold classification. To ensure the number of selected features is minimal while the most discriminant information is preserved, we propose a filtering method based on average information gain. For protein fold classification, we follow the hierarchical scheme in the previous study. That is, a sequence is first classified into its protein class and then into the fold. In order to take advantage of the diverse information from distinct feature sets, we adopt some well-known methods for information fusion. The experimental results reveal that our method achieves recognition rates of 88.3% and 75.7% for 4 protein classes and 27 folds, respectively, which outperform the previously best results 87% and 69.6%, respectively.**

***Index Terms*—Protein, Fold, SVM, HLA, Fusion.**

I. Introduction

In living organisms, each cell contains proteins that perform vital functions. In order to carry out these functions, proteins must fold into a three-dimensional structure. Hence, by analyzing the folding structures, we can understand protein functions. Since proteins are polymers of 20 amino acids joined by peptide bonds, the sequence of amino acids is termed as the primary structure.

As increasing number of proteins are produced by large-scale sequencing projects, to extract their structural information becomes an important issue. Among the topics in the bioinformatics field, the *fold recognition problem* is one of the subjects that have been intensively studied. That is, given a protein with its sequence of amino acids, the task is to decide which fold group the sequence belongs to.

In this paper, we re-examine the work of Ding and Dubchak [4], and then try to investigate how to achieve a high recognition rate. In their research, they utilized both *neural network* (NN) and *support vector machine* (SVM) to perform fold prediction. The experimental results reveal that the highest accuracy, 56.5%, is achieved by selecting 'CSHPV' from the set of entire 'CSHPVZ' features. From the statistical point of view, it turns out that relevant features are necessary to achieve a high recognition accuracy. Thus, by 2003, Huang *et al.* [5] proposed an MLP-based (multilayer perceptron) feature selection scheme to address the problem. Unfortunately, compared to the result by taking all features, the feature subset obtained from the MLP-based method does not achieve any improvement. In their study, however, they devised two new feature sets, bi-gram and spaced bi-gram, and proposed a *hierarchical learning architecture* (HLA). They demonstrated that both new feature sets and hierarchical schemes do facilitate protein and fold discrimination. The prediction accuracy of the first level (class classification) and overall prediction (fold recognition) accuracy are 84.4% and 65.5%, respectively.

In 2007, Lin *et al.* [8] proposed a feature selection and combination method for solving this problem. In their study, classifiers trained from distinct features subsets are combined according to a rank/score-based diversity. By means of selecting features carefully, they achieved 87% and 69.6% recognition accuracies for the protein class and fold

*This research work was partially supported by the National Science Council of Taiwan under contract NSC-97-2221-E-110-064.

†Corresponding author. E-mail:cbyang@cse.nsysu.edu.tw (Chang-Biau Yang).

C.-Y Hor and C.-B Yang are with the Department of Computer Science and Engineering, National Sun Yat-sen University, Kaohsiung, Taiwan. (email: cbyang@cse.nsysu.edu.tw)

S.-H Shiau is with the Department of Computer Aided Media Design, Chung Jung University, Tainan, Taiwan. (e-mail:shiaush@mail.cjcu.edu.tw)

classification, respectively.

Motivated by Lin's [8] approaches for feature subset selection, in this paper, we try to select each feature individually. Our feature selection is a filtering method which takes advantage of information gains. To preserve maximal discriminant information with minimal features selected, we first rank features according to their information gains. Then we average these gains and locate the feature at which maximal average gain drop occurs. Besides, we also perform some well-known classifier combination strategies. Our experimental results show that the best accuracies are 88.3% and 75.7% for the protein class and fold classification problem, respectively.

The rest of this paper is organized as follows. Section II introduces the basic material we use to perform the classification task, including the datasets, classifiers and feature evaluation method. Section III depicts our strategy that is devised to construct the entire architecture. In Section IV, the experimental results are demonstrated and then compared with the previous works. Finally, we give our conclusion in Section V.

II. Preliminaries

A. Data sets

The experimental data sets are obtained from Ding's website, including the training and testing sets. These data sets were retrieved from the *structural classification of proteins* (SCOP) [11]. According to the specification, any two proteins with aligned sequences longer than 80 residues are less than 35% and 40% identity within the training and testing data, respectively. No pair of sequences between the training and testing proteins are more than 35% identity. Table I lists the numbers of proteins and folds for the training and testing data. There are totally 311 and 383 sequences for training and testing, respectively. In the table, it is observed that sequences can be divided into four classes, α, β, α/β, and α+β. Thus, corresponding to these classes, there are 54, 109, 115 and 33 training sequences and 61, 117, 143, and 62 testing sequences.

TABLE I
SUMMARY OF TRAINING AND TESTING PROTEINS.

Class	Fold	No. of Training Proteins	No. of Testing Proteins
1.all-α	1.α_1	13	6
	2.α_2	7	9
	3.α_3	12	20
	4.α_4	7	8
	5.α_5	9	9
	6.α_6	6	9
2.all-β	7.β_1	30	44
	8.β_2	9	12
	9.β_3	16	13
	10.β_4	7	6
	11.β_5	8	8
	12.β_6	13	19
	13.β_7	8	4
	14.β_8	9	4
	15.β_9	9	7
3.α/β	16.$(\alpha/\beta)_1$	29	48
	17.$(\alpha/\beta)_2$	11	12
	18.$(\alpha/\beta)_3$	11	13
	19.$(\alpha/\beta)_4$	13	27
	20.$(\alpha/\beta)_5$	10	12
	21.$(\alpha/\beta)_6$	9	8
	22.$(\alpha/\beta)_7$	10	12
	23.$(\alpha/\beta)_8$	11	7
	24.$(\alpha/\beta)_9$	11	4
4.α+β	25.$(\alpha+\beta)_1$	7	8
	26.$(\alpha+\beta)_2$	13	27
	27.$(\alpha+\beta)_3$	13	27
Total		311	383

B. Support Vector Machines

Support vector machine (SVM) [15], [3] is a well-established technique for data classification. Given a training set of n-dimensional instances and label pairs $(\mathbf{x_i}, y_i), i = 1, 2, \ldots, N$ where $\mathbf{x_i} \in \mathbf{R^n}$ and $y \in \{-1, +1\}$, the SVM solves the following optimization problem:

$$\begin{aligned} &\min_{\mathbf{w},b,\xi_\mathbf{i}} \frac{1}{2}\mathbf{w^T w} + C\sum_{i=1}^{N}\xi_i, \\ &subject\ to\ y_i(\mathbf{w^T}\phi(\mathbf{x}_i) + b) \geq 1 - \xi_i, \\ &\xi_i \geq 0. \end{aligned} \tag{1}$$

The function ϕ maps the training vectors $\mathbf{x}_i$ into a higher dimensional space, namely feature space. SVM finds a linear separating hyperplane with normal vector $\mathbf{w}$ and offset b that constitutes the maximal margin in the feature space. The slack variable ξ_i is introduced to ensure that feasible solutions always exist. C denotes the penalty of errors in the optimization problem. To describe the similarity between vectors in the feature space, the kernel function, $K(\mathbf{x}_i, \mathbf{x}_j) \equiv \phi(\mathbf{x}_i)^T\phi(\mathbf{x}_j)$, is defined. The kernel function also determines the complexity of the target decision boundary. Among

kernel functions, we adopt *radial basis function* (RBF) as it yields best results.

C. Information Gain

In information theory, *entropy* is a measurement of the uncertainty that is associated with a random variable [1]. The entropy H of a discrete random variable X with possible values $x_1, x_2, \ldots, x_h$ is formulated as

$$H(X) = \sum_{i=1}^{h} p(x_i) \log p(x_i), \quad (2)$$

where $p(x_i)$ denotes the probability that variable X is of value x_i.

Information gain $I(X, Y)$ quantifies the dependence between the joint distribution of X and Y, and it is defined as

$$I(X, Y) = H(X) + H(Y) - H(X, Y), \quad (3)$$

where $H(X, Y)$ is the joint entropy of X and Y. If we associate X and Y with the feature and class label, information gain can be regarded as the relevance measure between these two items. When X and Y are statistically independent, $I(X, Y) = 0$. As for the feature selection, information gain is capable of accounting for the feature relevance with respect to the class label. The higher the gain is, the more relevant a feature is.

D. Hierarchical Learning Architecture(HLA)

By observing the data sets, we find that the folds are separated into 4 main protein classes, α, β, α/β, and $\alpha+\beta$. Within each class, it can be further categorized into several different folds. According to the characteristic of the data sets, Huang *et al.* [5] proposed a two-level classifier configuration, called *hierarchical learning architecture* (HLA). Given an unknown protein, it is recognized which class it belongs to in the first level. Then, a second-level classifier performs fold recognition. The configuration totally yields 5 classifiers involved for fold recognition. The entire architecture is illustrated in Fig. 1.

Huang compared the performance between HLA and non-HLA methods. Their experimental results show that HLA ensemble does outperform non-HLA one with the same feature sets. Although HLA ensemble can achieve higher recognition rates than non-HLA ensemble, the main drawback is: proteins that are incorrectly classified in the first level can not be recovered in the subsequent level.

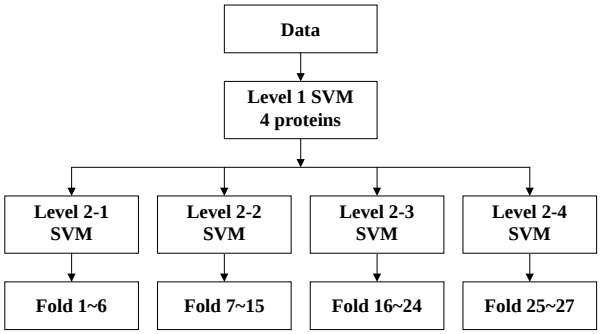

Fig. 1. The hierarchical learning architecture.

E. Features

Ding and Dubchak [4] proposed 125 features that are categorized into 6 groups. These feature sets are amino acid composition (C), predicted secondary structure (S), hydrophobicity (H), normalized van der Waals volume (V), polarity (P), and polarizability (Z). We obtain the data sets from their website. We leave these features intact except normalization.

To increase discriminant information for classification, we follow Huang's instruction [5] to construct bi-gram (B) and spaced bi-gram (SB) feature sets. Each of these two feature sets contains the combinations of any two from 20 common amino acids. Consequently, there are 400 features for each feature set, which is relatively large compared to the lengths of all experimental sequences.

Since features in B and SB sets are quite sparse and the number of features is quite large, we first perform *principal component analysis* (PCA) [9], [12] to these two feature sets in each level of the training data. Then we transform the testing data to the corresponding principal coordinates. To be explicit, let the matrix of the training data is of dimension $n \times N$, where n and N denote the numbers of the original features and training data, respectively. Since the maximal effective rank of the data matrix is $min(n, N)$, this implies that the dimension of non-null space spaned by the matrix should not exceed this number. Hence, for both of transformed B and SB features in the first level, we obtain 311 $(min(400, 311))$ features for each set, which corresponds to the size of the training data set. Identically, in the second level of process,

TABLE II
SUMMARY OF FEATURE SETS.

Symbol	Feature	Dimension
C	Amino acids composition	20
S	Predicted secondary structure	21
H	Hydrophobicity	21
P	Polarity	21
V	Normalized van der Waals volume	21
Z	Polarizability	21
A	PCA of B feature set	311/54/109/115/33
B	PCA of SB feature set	311/54/109/115/33

the extracted feature numbers are given as 54, 109, 115, and 33 for each feature set. PCA helps us to use fewer features to represent the information. Thus concatenating with the above mentioned 125 features, it yields 747, 233, 343, 355 and 191 features for further selection. Table II summarizes all available features used in this paper. Each node in the HLA ensemble owns its individually available features.

III. PROTEIN FOLD CLASSIFICATION

A. Feature Selection and Combination

According to Huang's results, they pointed out that training involving 'CSHPVZ' features obtains a higher accuracy than that involving only 'B' (or 'SB') feature set. This may imply that 'CSHPVZ' feature set are more significant than 'B' (or 'SB'). Besides, they also showed that combining 'B' and 'SB' with 'CSHPVZ' features further improves the accuracy. It implies that 'B' and 'SB' feature sets do contain some extra competent information over 'CSHPVZ' one. This motivates us to treat 'CSHPVZ' and 'B+SB' feature sets separately.

Since a lot of available features can be used, this raises another issue: how many features should be preserved? Although keeping only a subset of features would simplify the behavior represented by the classifier, it usually would deteriorate the fitness to the data. Even though, we would probably benefit from achieving a better generalization. Thus, it seems reasonable to impose some constraints to the learning algorithm [10]. Given the same kernel function in the SVM training, the classifier's complexity may be determined by the number of involved features.

In the feature selection stage, our idea to keep features is as follows: to preserve maximal discriminant information with the minimal number of features selected. Thus, we adopt the idea of average information gain as it represents the tradeoff between the simplicity and goodness of fitness of the model. Our procedure for feature selection is given as follows:

Procedure Feature selection.
Input: Training data with feature set S, where $|S| = n$ and S may be 'CSHPVZ' (125 features) or 'AB' (108, 218, 230, or 66 features).
Output: Selected features, a subset of S.
Step 1: Calculate the information gain of each feature.
Step 2: Sort the features according to their information gains in nonincreasing order and obtain $f_1,f_2,\ldots,f_n$.
Step 3: For each p, $1 \le p \le n$, calculate the average information gain as follows:
$F_p = \frac{1}{p}\sum_{1\le j\le p} f_j$.
Set smoothing factor $t = 1$.
Step 4: For $k1 = t$, $k2 = t + 1$, find the feature at which local maximal drops of average information gains occur according to:

$$p1 = arg \max_{1\le p\le n} \frac{F_p - F_{p+k1}}{F_{p-k1} - F_p}. \quad (4)$$

$$p2 = arg \max_{1\le p\le n} \frac{F_p - F_{p+k2}}{F_{p-k2} - F_p}. \quad (5)$$

Step 5: Check if selected feature subsets $D(k1) = [f_1,f_2,\ldots,f_{p1}]$ and $D(k2) = [f_1,f_2,\ldots,f_{p2}]$ are identical. If so, output $D(k1)$ and then stop.
Step 6: Build classifiers and use the training data to validate if $D(k2)$ outperforms $D(k1)$. If not, output $D(k1)$ and then stop. If $D(k2)$ is superior and achieves 100% classification rate for the training data, output $D(k2)$ and then stop. Otherwise, set $t = t + 1$ and goto Step 4.

k is the smoothing factor for the calculation of the maximal drop and p serves as the cut-out point for feature selection. To determine the suitable value of k, we adopt a greedy approach to evaluate the preserved features. That is, once an appropriate k is obtained, the iteration is stopped. In this paper, we find that $k = 1$ or 2 is capable of picking out maximal drops and thus it is used through out our experiments.

B. Majority Vote

The *majority vote* (MAJ) [14] assigns an unknown input **x** to the most representative class

among classifiers' outputs. Given each classifier' output as an m-dimensional binary vector $(d_{i,1}, d_{i,2}, \ldots, d_{i,m}) \in \{0,1\}^m, 1 \leq i \leq L$, where $d_{i,j} = 1$ if $\mathbf{x}$ is labelled as class j by the classifier i, otherwise $d_{i,j} = 0$. The majority vote picks up class c if

$$c = \max_{1 \leq j \leq m} \sum_{i=1}^{L} d_{i,j}. \tag{6}$$

For example, given a system consisting of 2 pairwise-SVMs and $m = 3$, each SVM will output $(d_{i,1}, d_{i,2}), (d_{i,1}, d_{i,3}), (d_{i,2}, d_{i,3})$ where $i = 1, 2$ and $d_{i,j} \in \{0, 1\}$. The majority vote approach manages to choose the plurality among these outputs.

If classifiers' predictions are not equally accurate, then it would make sense to assign the more competent member more power in making the final decision. This constitutes the idea of *weighted majority vote* (WMAJ).

$$c = \max_{1 \leq j \leq m} \sum_{i=1}^{L} b_i * d_{i,j}, \tag{7}$$

where b_i represents the weight of each classifier in the ensemble. According to the result proposed by Kuncheva [6] , if L *independent* classifiers, with individual accuracies as $p_1, \ldots, p_L$, are involved to solve a two-class recognition problem, to maximize the overall accuracy, the weight b_i for each member should be

$$b_i \propto \log \frac{p_i}{1 - p_i}. \tag{8}$$

This optimal weight emphasizes the importance of each classifier based on its accuracy and it is adopted as a standard choice for the well-known Adaboost algorithm [6].

C. Behavior Knowledge Space

Behavior knowledge space (BKS) [13] is a table look-up approach for classifier combination. Let us consider a classification task for m classes. Assume we have totally L classifiers composing an ensemble and these classifiers collaborate to perform recognition. Given an input $\mathbf{x}$, the ensemble produces a discrete vector $E(\mathbf{x}) = (d_1, \ldots, d_L)$ where each $d_i \in \{1, \ldots, m\}$ represents the output of the ith classifier. Thus, the number of possible combinations of L classifiers' outputs is m^L. For the entire training set, the ensemble's outputs constitute an intermediate knowledge space, which characterizes all classifiers' preferences.

TABLE III
AN EXAMPLE OF THE BKS TABLE.

Prediction		True class			
D1	D2	P1	P2	P3	P4
P1	P1	**10**	3	3	0
P1	P2	3	0	**6**	1
P1	P3	**5**	4	0	2
⋮	⋮	⋮	⋮	⋮	⋮
P4	P3	0	2	2	**5**
P4	P4	0	0	1	**6**

In practice, the algorithm can be implemented by a look-up table, called BKS table. Each entry in the table contains L cells where each cell accumulates the number of the true classes of training inputs falling in. During the recognition stage, the ensemble first collects each classifier's output $D_i(\mathbf{x}), i = 1, \ldots, L$. Then it locates which entry matchs the outpus, and then picks up the class label corresponding to the plurality cell.

Table III illustrates a BKS table of HLA level 1 with $m = 4$ and $L = 2$. $D1$ and $D2$ represents two classifiers. Entries below $D1$ and $D2$ are all possibile predictions. Cells below 'True class', which are 'P1', 'P2', 'P3', and 'P4', are numbers of class labels that the training data elements fall into. Thus, each entry in the table contains the most representative labels that associate with the perferences of the ensemble.

As mentioned in Section II, the drawback of HLA ensemble is that if a protein is incorrectly classified in the first level, then it can not be recovered in the second level. Hence, we use the *hierarchical behavior knowledge space* (HBKS) [2] to remedy the problem. Instead of processing the data sequentailly by only one classifier in each level, the HBKS ensemble forces all its members to classify simultaneously. Hence, five votes are involved to make decisions. Since the information is combined from different levels, we can think HBKS perform information fusion in a vertical manner. Table IV shows an example of the HBKS table. Because the entries contain the perferences of both levels, the ensemble may fully utilize the the information to make final decision.

TABLE IV
AN EXAMPLE OF THE HBKS TABLE.

Level 1 Prediction	Level 2 Prediction				True class
D1 P1-P4	D2 F1-F6	D3 F7-F15	D4 F16-F24	D5 F25-F27	F1-F27
⋮	⋮	⋮	⋮	⋮	⋮

TABLE V
SELECTED FEATURES AND CLASSIFICATION ACCURACIES (%)ON THE TRAINING DATA.

	k=1		k=2	
Feature combination	(a)	(a+b)	(a)	(a+b)
Level 1	95.5(38)	96.4(64)	95.5(38)	96.4(64)
Level 2-1	46.3(2)	59.2(10)	100(22)	100(30)
Level 2-2	100(58)	100(85)	100(58)	100(85)
Level 2-3	100(53)	100(74)	100(53)	100(74)
Level 2-4	100(23)	100(28)	100(23)	100(28)

(a): The subset of the 'CSHPVZ' feature set.
(a+b): The subset of the 'CSHPVZ'+'AB' feature set.

IV. EXPERIMENTAL RESULTS

A. Feature Selection

We use the training data to compute the information gains in the first level. Then we arrange the training data according to their protein classes and perform the same procedure in the second level. By applying the approach mentioned in Section II, we select features for the HLA-SVM trainings. The smoothing factors, preserved feature numbers (shown in parenthesis) and their corresponding accuracies (in percentage) on the training data are listed in Table V. In the table, (a) and $(a+b)$ denotes the subsets of 'CSHPVZ' and 'CSHPVZ'+'AB' feature sets, respectively. It is observed that except for the first protein class in level 2, classification accuracies are generally not affected by smoothing factors. Besides, for this special case, the accuracy is significantly low when $k = 1$. Thus, $k = 2$ is adopted for the subsequent experiments. All selected features are listed in Table VI. In this table, the selected features are arranged in hierarchy. All entries are futher divided into two parts, which represent (a) and $(a+b)$ feature sets. Each feature in the table is expressed by concatenating its feature symbol, as illustrated in Table II, and feature identifier. For example, $C2$ denotes the second feature from the amino acids composition set.

TABLE VI
SELECTED FEATURES.

	Selected features
Level 1	C2,C3,C5,C6,C8,C9,C14,C15,C18,S1, S2,S3,S6,S7,S9,S12,S16,H3,H4,H5,H7, H8,H9,H13,H14,H16,H17,H18,H19,V1, V4,V9,V13,P2,P7,P19,Z5,Z18
	A1,A5,A24,A26,A53,A55,A61,A64, A78,A81,A90,A110,A116,A117,A162, A179,A192,B10,B14,B43,B48,B60,B90, B127,B130,B140
Level 2-1	C1,C2,C3,C4,C5,C6,C9,C15,C16,C17, S1,S2,S3,H4,H11,H12,V2,V7,V8,V12, Z12,Z21
	A8,A13,A21,A22,A38,B22,B36,B38
Level 2-2	C2,C4,C5,C6,C7,C10,C11,C12,C13,C14, C17,C19,S4,S5,S6,S11,S14,S19,S20,H2, H3,H5,H6,H11,H15,H16,H17,H18,H19, H20,H21,V1,V2,V3,V4,V10,V12,V17, V18,P1,P6,P8,P9,P11,P13,P14,P17,P20, P21,Z3,Z4,Z5,Z7,Z11,Z12,Z14,Z16,Z18
	A1,A6,A8,A20,A38,A41,A47,A48,A51, A52,A56,A71,A87,A101,B4,B5,B28, B36,B43,B45,B46,B48,B62,B65,B95, B100,B102
Level 2-3	C3,C5,C6,C7,C8,C9,C11,C12,C14,C15, C16,C17,C20,S3,S5,S6,S7,S8,S11,S12, S14,S15,H3,H5,V1,V4,V13,V21,P2,P6, P7,P8,P9,P10,P12,P13,P14,P19,P20,Z2, Z3,Z5,Z6,Z8,Z10,Z11,Z13,Z15,Z16,Z17, Z18,Z19,Z20
	A1,A9,A34,A58,A59,A68,A89,A92, A106,B8,B9,B12,B29,B41,B45,B59, B69,B77,B81,B86,B88
Level 2-4	C1,C2,C3,C7,C8,C9,C10,C13,C14,C16, S3,S14,S17,S19,V2,P3,P4,P5,Z11,Z17, Z18,Z19,Z20
	A1,B1,B7,B9,B28

B. Classification Accuracies

In this subsection, we shall focus on classification accuracies, including HLA and HLA-fusion. The accuracy is defined as $Q = \sum p_i/N$, where p_i denotes the number of testing targets that belong to class i and are correctly classified, and N denotes the total number of testing targets. Table VII shows accuracies of all configurations for the testing data. The (a) and (a+b) are HLA-SVMs trained with 'CSHPVZ' and 'CSHPVZ'+'AB' feature subsets. During the recognition stage, each testing sequence is first classified by the level 1 classifier. The level 1 classifier then determines which classifier in the level 2 should carry out the subsequent classification task. Finally, the designate level 2 classifier makes the fold recognition. It shows that training with partial features achieves better performance than that with full features. In addition, the improvement in level 2 are more significant than that in level 1. This suggests that feature selection imposes more effects on the data with small amount. If we compare

the accuracies of HBKS and HLA SVM, we find that the accuracy of HBKS ensemble is superior to HLA with the same feature set. Hence, combining the classification information of both levels may occasionally achieves better performance.

If we consider the last two configurations, accuracies of (a+b) are generally higher than those of (a), except for the fourth class of level 2. This supports Huang's conclusion that 'AB' (or 'S'+'SB') feature set does have its unique competence over 'CSHPVZ' one. In our study, features are selected individually rather than by group according to their degrees of relevance. Hence, the truly influential features in each group are not easily overlooked even if most of other features in the same group are not so prominent. Beyond that, the learning algorithm may be able to pay more attention on relevant features and thus enhance its generalization capability.

In the previous works [7], [8], they demonstrated the connections between diversity and accuracy in some cases. We also try to investigate the possibility for the recognition improvement. We perform three classifier fusion methods: pairwise majority vote, weighted majority vote and behavior knowledge space. The first one gives each classifier equal weight to make decision and is commonly used in our daily electoral events. Instead of assigning 1,2,3,...,c rank scores as outputs in the rank/score fusion, the pairwise majority vote only gives 0 and 1 for each pair of loser and winner and thus is more straightforward. The second one gives superior classifiers more power to vote. The better the classifier performs on the training data, the higher weight it receives for its subsequent prediction. We assume that the two candidate classifiers are independent and assign them weights, $\log(p_i/(1-p_i))$, according the optimal weight rule. The results are shown in Table VIII

Although the above two methods work reasonably well on the alternative faulty cases, the drawback is that they are easily trapped into the situation of double failures. In contrast, BKS is able to handle this kind of case because it can memorize each classifier's preferences of the training data, including both correct and incorrect predictions. Therefore, this mechanism provides the BKS ensemble the ability to amend the final decisions. The experimental results, illustrated in Table VIII, reveal that BKS outperforms other methods.

TABLE VII
THE ACCURACIES OF HLA CLASSIFICATION.

Method	HLA NN CSHPVZ[1]	HLA SVM CSHPVZ +S+SB[1]	HLA SVM (a)	HBKS SVM (a)	HLA SVM (a+b)	HBKS SVM (a+b)
Accuracy of Level 1(%)	81.3	84.4	84.6	-	**86.9**	-
Accuracy of Level 2(%)	67.2	73.8	80.3	-	**81.9**	-
	52.1	60.7	62.4	-	**70.9**	-
	58.6	65.5	69.2	-	**76.9**	-
	48.4	58.1	**59.7**	-	58.1	-
Overall Accuracy(%)	56.4	64.2	67.4	67.4	72.8	**74.2**

1: Data from Huang *et al.* [5].
(a): The subset of the 'CSHPVZ' feature set.
(a+b): The subset of the 'CSHPVZ'+'AB' feature set.

TABLE VIII
ACCURACIES OF CLASSIFIER FUSION.

Method	HLA NN RANK[1]	(a)+(a+b) MAJ	(a)+(a+b) WMAJ	(a)+(a+b) BKS
Accuracy of Level 1(%)	87	84.9	86.9	**88.3**
Accuracy of Level 2(%)	69.6	68.1	69.5	**75.5**

1: Rank/score fusion, data from Lin *et al.* [8].
(a): The subset of the 'CSHPVZ' feature set.
(a+b): The subset of the 'CSHPVZ'+'AB' feature set.

Let us consider conventional HLA BKS as horizontal because the information fusion only takes place in the same level . On the contrary, HBKS combines information in a vertical manner. That is, the information are fused in different levels. As mentioned above, HLA simplifies the learning process while sometimes leads to early fault. HBKS may avoid the shortcoming while the improvement is not guaranteed. This inspires us to aggregate the information of level 2 from these two configurations. Table IX illustrates the results for information fusion. Although the improvements are not significant, combinations in both horizontal and vertical schemes do slightly facilitate the discrimination. The best result is 75.7%.

V. Conclusion

In this paper, we investigate a feature selection and combination method for protein fold classification. For feature selection, we propose a filtering method based on average information gain. We then use the selected features to construct the HLA SVM. In order to fully utilize the information from distinct feature sets, we perform three methods for classifier

TABLE IX
ACCURACIES OF COMBINING HORIZONTAL AND VERTICAL ENSEMBLES OF BKS FUSION.

Fusion combination	Accuracies(%)
110	**75.7**
101	**75.7**
011	74.4
111	**75.7**

001: HBKS SVM (a).
010: HBKS SVM (a+b).
100: (a)+(a+b) BKS.

fusion. The experimental results exhibit that our results are better than those from the previous studies, in which the recognition rates for 4 protein classes and 27 folds are 88.3% and 75.7%, respectively. From the results of level 2, we find that accuracies of second and fourth protein classes are relatively low. To improve this, there should be additional discriminant features to be explored. Although HLA ensemble can achieve higher recognition rates than non-HLA ensemble, proteins that are incorrectly classified in the first level can not be recovered. Hence, we try to use HBKS to address the issue. However, the performance improvement is not significant. We believe that a mixture of HLA and non-HLA would be more appropriate. Besides, since the feature selection is a combinatorial problem, an approach between one-by-one and group-by-group may be a better strategy. Our future research would try to resolve the above issues.

REFERENCES

[1] D. J. C. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, 2003.

[2] H. Cecotti and A. Belaïd. Hierarchical behavior knowledge space. In *MCS*, pages 421–430, 2007.

[3] C. C. Chang and C. J. Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at http://www.csie.ntu.edu.tw/∼cjlin/libsvm.

[4] C. H. Q. Ding and I. Dubchak. Multi-class protein fold recognition using support vector machines and neural networks. *Bioinformatics*, 17:349–358, 2001.

[5] C.-D. Huang, C.-T. Lin, and N. R. Pal. Hierarchical learning architecture with automatic feature selection for multiclass protein fold classification. *IEEE Transaction on nanobioscience*, 2(4):221–232, 2003.

[6] L. Kuncheva. "fuzzy" versus "nonfuzzy" in combining classifiers designed by boosting. *IEEE Transactions on Fuzzy Systems*, 11(6):729– 741, 2003.

[7] L. Kuncheva. Measures of diversity in classifi er ensembles and their relationship with the ensemble accuracy. *Machine Learning*, 51(2):181– 207, 2003.

[8] C. Y. Lin, K.-L. Lin, C.-D. Huang, H.-M. Chang, C. Y. Yang, C.-T. Lin, C. Y. Tang, and D. F. Hsu. Feature selection and combination criteria for improving predictive accuracy in protein structure classification. *IEEE Transaction on nanobioscience*, 6(2):186–196, 2007.

[9] K. V. Mardia, J. T. Kent, and J. M. Bibby. *Multivariate Analysis*. Academic Press, 1979.

[10] T. M. Mitchell. *Machine Learning*. McGraw Hill, 1997.

[11] A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *Journal of Molecular Biology*, 247:536 – 540, 1995.

[12] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2008. ISBN 3-900051-07-0.

[13] S. Raudys and F. Roli. The behavior knowledge space fusion method: Analysis of generalization error and strategies for performance improvement. In *In Proc. Int. Workshop on Multiple Classifier Systems (LNCS 2709*, pages 55–64. Springer, 2003.

[14] D. Ruta and B. Gabrys. A theoretical analysis of the limits of majority voting errors for multiple classifier systems. *Pattern Analysis & Applications*, 5(4):333–350, 2002.

[15] V. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag, 1995.