Multi-Class Protein Fold Recognition Using Multi-Objective Evolutionary Algorithms

Stanley Y. M. Shi, P. N. Suganthan, *Senior Member IEEE* and Kalyanmoy Deb, *Senior Member IEEE*

KanGAL Report Number 2004007

Abstract— **Protein fold recognition (PFR) is an important approach to structure discovery without relying on sequence similarity. In the pattern recognition terminology, PFR is a multi-class classification problem to be solved by employing feature analysis and pattern classification techniques. This paper reformulates PFR into a multi-objective optimization problem [7] and proposes a Multi-Objective Feature Analysis and Selection Algorithm (MOFASA). We use support vector machines as the classifier. Experimental results on the Structural Classification of Protein (SCOP) data set indicate that MOFASA is capable of achieving comparable performances to the results reported in [10]. In addition, MOFASA identifies relevant features for further biological analysis.**

Index Terms— **protein fold recognition, feature selection, multi-class classification, multi-objective evolutionary algorithm, NSGA-II, support vector machines.**

I. INTRODUCTION

RECENT structural genomic initiatives and improvements in experimental methodologies have populated the biological databases at a rapid pace. Current release of the protein data bank (Jun 2004) [12] contains more than 26,000 determined protein 3D structures. However, this is far beyond satisfactory in comparison to the more than 280,000 known protein sequences (May 2004) [6]. As experimental methods are time consuming and expensive [2], it is important to develop accurate automatic structure prediction algorithms to determine the structure of new sequences.

Most of the computational protein structure prediction approaches are based on sequence similarities [3]. If a new sequence has a high sequence similarity to a protein with known structure, the new protein may belongs to a similar fold and share the common evolutionary ancestor with this known protein [5]. In this situation, sensitive sequence comparison methods might be applicable for detecting the close evolutionary relationships between proteins [13]. However, these methods are not efficient when two proteins have a closely related structures with no obvious similarities between sequences. Protein fold recognition can be used to detect such kinds of relationship [17].

Ding and Dubchak [10] recently proposed taxonometric approach to determining structural similarity without relying on sequence similarity. They applied the machine learning methods such as neural networks and support vector machines (SVM) in multi-fold protein recognition. Both one-versus-all (OVA) and one-versus-one (OVO) classification strategies have been studied. But the accuracy is relatively low. Motivated by Ding's work, Tan $et\ al$ [2] devised an ensemble classifier, eKISS, for K-fold protein classification. eKISS first generates a group sophisticated rule based classifiers as base classifiers, then combines all these classifiers into a new ensemble classifier. Among these classifiers, (K) classifiers are derived from OVA and $(K(K-1)/2)$ classifiers are derived from OVO. The ensemble classifier is reported [2] to be robust and generates a set of helpful rules for further biological analysis. Their result is difficult to compare with the original Ding's work for only folds with more than eight samples were used in the experimental analysis in [2].

In this paper, we study the feature selection problem in the context of multi-class PFR. We generalize the wrapper method [19] by using both training and testing accuracy to guide the feature subset selection. We formulate the feature subset selection problem into a three objective optimization problem similar to an earlier study [8] and propose a Multi-Objective Evolutionary Algorithm (MOEA) to solve it. Since SVM are strong classifiers with good generalization capability and OVO strategy is known to be effective for multi-class recognition [10] [14], we embed the OVO SVM as our classifier. We call our method multi-objective feature analysis and selection algorithm (MOFASA).

The organization of this paper is as follows. In section II, we provide the background of feature analysis and classification problems. In section III, we describe the multi-objective evolutionary algorithm for optimizations. This is followed by section IV, the methodology and experimental results of MOFASA, Finally, in section V, we discuss the performance of the proposed algorithm for PFR and suggest some future research directions.

II. FEATURE SELECTION AND CLASSIFICATION

A. Feature Selection

The protein fold recognition is a classical multi-class classification problem involving feature analysis. Feature selection or feature extraction operation can be used

Stanley Y. M. Shi is from Nanyang Technological University;
Dr. P. N. Suganthan is from Nanyang Technological University;
Dr. Kalyanmoy Deb is from Indian Institute of Technology, Kanpur.

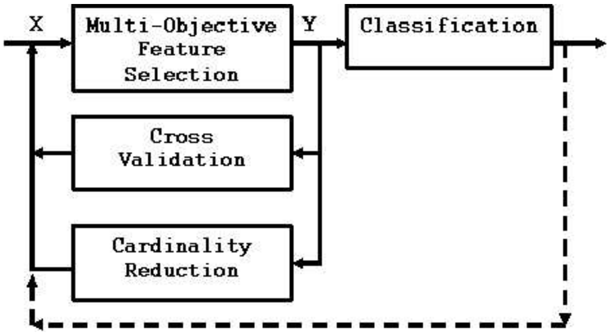

Fig. 1. Extended wrapper method with an additional testing feedback loop. X: the input features, Y: the selected features.

to perform feature analysis. In biological classification problems, it is important to be able to explain the reasoning behind the decisions made. Hence, the feature analysis task should ideally be performed by feature subset selection as feature extraction operation does not allow us to easily relate the decisions made to original observations.

The feature selection problem can be defined as follows: given a set of candidate features and a collection of data samples, select a subset that performs the best according to specified criteria. The selected feature subsets will not only reduce the computational requirements, but also likely to achieve a better performance due to the finite sample size effects [16] [20]. Further, the selected combinations of features may facilitate biological analysis leading to further understanding and insight into corresponding biological functions.

The feature selection can be categorized into embedded, filter and wrapper approaches [4]. The embedded approach may provide some useful rules for classification while it may sacrifice accuracy as it combines feature selection and classification into one process. Filter approach is faster but comparatively lower in accuracy, since feature subsets selection process is independent of the classification. The wrapper approach [19] uses two processes–one searches for a feature subset, the other evaluates the selected feature subset using the same classification algorithm. This two-step process is repeated and directed by the classification performance feedback until an optimal feature subset is obtained. In other words, the feature subset is optimized for the classifier. While feature subset evaluations require higher computing resources, this method is more likely to yield higher accuracy. The traditional wrapper method uses only validation results to evaluate a feature subset. In MOFASA, we extend the wrapper concept by using an additional feedback loop shown by dashed lines in Figure 1, to integrate testing accuracy too. In the multi-objective optimization discussed in section 3, this loop will generate trade off solutions in the Pareto front, from which we can observe the important features.

B. SVM for Classification

Support Vector Machines, derived from statistical learning theory, were originally designed for binary classification. Some favorable characteristics of the SVM are the absence of local minima and the implicit kernel mapping scheme from the input feature space to a highly non-linear feature space. The basic idea can be outlined as follows: First, the input vectors are implicitly mapped into a higher dimensional feature space (possible with a higher dimension) using the kernel trick [24]. Then seek an optimized linear decision boundary in the feature space which is able to separate two classes with least error and maximal margin.

SVM has been successfully used in protein analysis [25] and demonstrated high classification accuracy and good generalization performance. In Ding and Dubchak's work [10], SVM performs better than neural networks. In our early work [23], we noticed that SVM is more accurate than weighted *k* nearest neighbors classifiers. The rule based decision tree methods in general suffers from the relatively higher variance and bias. In other words, different tree solutions may vary greatly and the training errors may be considerably less than the test errors [15]. Based on these observations, we choose LibSVM [21] as the classifier. Since SVM is naturally suited for binary classification scenarios, extending it for multi-class classification problems is an on-going research problem. Most of the current work is based on different classifiers combination strategy. In [14], three most popular strategies, OVO, OVA and Directed Acyclic Graph (DAG) SVM were compared and OVO and DAG SVM were found to perform better. In this paper, we will employ the OVO strategy.

III. Multi-Objective Evolutionary Algorithms(MOEAs)

A multi-objective optimization problem has a number of objective functions which are to be minimized or maximized [7]. MOEAs use evolutionary algorithms to optimize these objectives simultaneously and give out a Pareto-optimal front (a solution set) for higher level analysis and decision. Many of the real world problem are multi-objective problem. MOEAs provide a insight to deal with these problems and therefore are the most active research directions in current optimization area.

As mentioned before, we formulate the PIR into three objectives. To handle the three objectives optimization problem, we use a fast elitist MOEA−Non-dominated Sorting Genetic Algorithm II (NSGA-II) [18][1]. We briefly describe the algorithm here: First, an offspring population is created from the same size parent population, by using genetic operators (such as single-point crossover and bit-wise mutation operators [11]). Then the two populations are combined together as

[1] The IEEE TEC paper describing NSGA-II for multi-objective optimization is judged as the FAST-BREAKING PAPER IN ENGINEERING by Web of Science (ESI) in February 2004.

a double-size mating pool. The duplicated solutions are removed from the mating pool. A *non-dominated* sorting [7] is applied to all the remaining solutions. After that, the new parent set was created by choosing solutions of different *non-dominated* fronts according the ascending order in the mating pool one by one. There are some different situations during this procedure. If the non-duplicate solution (here we can accept them as feasible solutions) is more than the slots available in the new parent population, a crowding distance [7] based selection is used to select the solutions that will make the diversity of the solutions maximum.

It is important to realize that in the classification problem, there may exist multiple classifiers resulting in an identical set of objective values. For example, the same size of selected features and same classification accuracy. We call these multi-model solutions in same group. To deal with this situation, a modification on the original NSGA-II was proposed in [8], which enables NSGA-II to deal with this situation and find multiple classifiers simultaneously. The key modification is as follows: When the number of distinct objective solutions in the last Pareto front n_l is smaller than the number of empty slots in the new parent population, these n_l distinct objective solutions will be first selected. The remaining slots will be filled with certain number of randomly selected multi-model solutions. The numbers of selected multi-model solutions for each group are proportionate the number of their appearance in the last front. Finally, new parent population created when all empty slots are filled.

NSGA-II has the following important features:

1) It uses an elitist principle so that best solution(s) in the previous iteration is (are) always included in the current iteration.
2) It uses an explicit diversity preserving mechanism, thereby allowing maintaining multiple trade-off solutions in an iteration
3) It emphasizes the on-dominated solutions, thereby ensuring convergence close to the Pareto-optimal solutions.

IV. Multi-Objective Feature Analysis and Selection Algorithm (MOFASA)

Protein fold recognition is a classification problem. Our objective is to provide more information, such as the relevant feature subsets, classification accuracy, and bias characters of both training and testing data sets. This information may assist the subsequent biological analysis and further experimental designs. As discussed before, we extend the wrapper approach by introducing an testing accuracy feedback loop. This loop expands the Pareto-optimal front to include trade-off solutions based on testing data set. Another loop is the traditional 10 fold cross validation accuracy loop. This loop expands the Pareto-optimal front based on the training data set. We will compare these results to analyze the differences between the training and testing sets. The third loop is the cardinality loop. As the name suggests, the cardinality loop will guide the NSGA-II to find more solutions with cardinality as small as possible. Thus, we formulate the PFR with three objectives, along similar lines to an earlier study [8]:

1) Maximize the cross validation accuracy on the training set.
2) Maximize the classification accuracy on the testing set.
3) Minimize the cardinality of feature subsets.

We use an l-bit binary string where l is the cardinality of the full feature set, to represent a solution. In a particular string, the features corresponding to the positions marked by 1 are selected. For example, in a 10-bit string, (1010010000), first, third and sixth features are chosen into the feature subset, and we use these features to induce and test classifiers. We use the OVO strategy for multi-class classification. The fold prediction is performed by the majority voting of $K(K-1)/2$ SVM classifiers' decisions where K is the number of folds. We calculate the 10-fold cross validation accuracy and test accuracy. We initialize each population member by randomly choosing at most 50% of string positions represented by $1s$. We set the population size to 100 and number of generations to 200. In the SVM, we use 4^{th} order polynomial kernel.

V. Experimental Evaluation

A. Data set

We use the same data set used in [10]. The training set was selected from the database built for 128-fold prediction problem. The database is based on the $PDB\text{-}select$ sets, each pair of proteins has no more than 35 percent of the sequence identity for the aligned subsequences longer than 80 residues. 27 most populated folds in the database with seven or more proteins are utilized. There are 313 protein samples in the this training set. PDB-40 set is used as an testing data set. The set contains the SCOP sequences having less than 40 percent identity with each other and 385 protein samples of the same 27 folds as in the training set. There are six groups of continuous features for each protein as summarized in Table 1. Each group of features is extracted independently. There are many duplicated (redundant) and irrelevant features among these 125 features. For example, the last 6 features in the van der Waar volume group are the same as the last 6 features in the polarizability group. In the predicted secondary structure group, the 4^{th} and 21^{st} features are irrelevant. Most of the values in the 4^{th} feature are zero and all the values of the 21^{st} feature are 100.

B. Analysis of Results

We use the standard Q-percentage accuracy [22] to handle both true positives and false positives. Q can be expressed as the sum of correctly predicted samples in each fold divided by total number of samples to be predicted. We calculated the average of cross validation and test Q-percent accuracy (AVQ) for each *non-dominated* solution in the last generation, and select the top 3 AVQ results as representatives.

We initially test each feature group separately. The top three AVQ results are presented in Table II. As we can see, MOFASA yields higher Q-percentage accuracy in all six feature subsets. The initial feature subset cardinality is set between 15 to 20(21). Intuitively, it is an approximation of backward elimination procedure. In Table II, we observe that different feature groups achieve different fold prediction accuracy. The amino acids composition feature subset performs the best, followed by predicted secondary structure, hydrophobicity, polarity, normalized van der Waals volume and finally polarizability.

There is useful information in the first *non-dominated* solution set, such as which features are useful in the fold recognition and the bias character of training and testing data sets. In figure 2, we plotted the frequency of each feature in the final feature subsets. The first two feature sets (C and S) have more high frequency features than the other groups. This may be the reason why they perform better in fold prediction as shown in Table II. We sorted the solutions according to the cross validation accuracy in Figure 3 and according to the testing accuracy in Figure 4. The x-axis and y-axis represent the accuracy rank and the feature groups, respectively. By comparing the two figures, we find that the protein secondary structure (S) feature group varies notably. Having more S features in the feature subsets increases the cross validation accuracy, while eliminating these features improves the testing accuracy. There are several features of this nature in the other sets too. These features may cause classifier bias on particular data sets. Further biological analysis of these features may reveal better understanding of their functions.

In multi-class fold recognition, prediction accuracy can be significantly affected by the number of samples in each fold. We combined eight folds with at least 13 samples into a new data set, and then tested the performance of the MOFASA. The results are presented in Table III. Again, MOFASA yields higher prediction accuracy.

Independent test can exam the generalization capability. We test MOFASA on independent set as well. Here the test set is divided into a new test set with 185 samples and a independent set with 200 samples. The training set remains unchanged. In addition to the six independent groups, we add an additional mixed group, which is composed of all the above six groups of features. The same procedures are employed to introduce the classifiers on these seven groups. Then these classifiers are tested on the independent set. The best three AVQ solutions are presented in Table IV together with the original work [10]. From these results we can see that MOFASA achieves comparable Q-percent prediction accuracy to the original study in all seven groups, which indicates a very good generalization capability.

TABLE I

Feature Groups

Group	Parameter	Dim
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

TABLE II

Independent Feature Group Test

Gp.	Cross Valid.(%)		Indep. Test (%)		Average (%)		Cd.
	MOF	Orig	MOF	Orig	MOF	Orig	
C	36.10	32.70	49.87	44.90	42.99	38.80	1
	34.50		49.87		42.19		1
	35.14		48.83		41.99		2
S	38.66	34.60	45.45	35.60	42.06	35.10	1
	40.58		43.12		41.85		2
	39.94		43.38		41.66		3
H	25.88	19.80	41.56	36.50	33.72	28.15	1
	25.88		41.30		33.59		2
	24.28		42.34		33.31		3
P	26.20	18.70	38.70	32.90	32.45	25.80	2
	24.92		38.96		31.94		3
	24.28		39.22		31.75		4
V	23.64	17.20	37.40	35.00	30.52	26.10	3
	24.60		35.84		30.22		4
	23.00		37.40		30.20		5
Z	22.04	14.60	37.14	32.90	29.59	23.75	1
	22.04		36.62		29.33		1
	25.88		32.47		29.17		2

$Gp.$: Group;
MOF: Multi-Objective Feature Analysis and Selection Algorithm;
$Orig$: Original Method in [10];
$Cd.$ Cardinality of feature subsets.

TABLE III

8 folds and 27 fold classification

Gp.	Cross Valid.(%)		Indep. Test (%)		Average (%)		Cd.
	MOF	Orig	MOF	Orig	MOF	Orig	
8 folds	90.10	62.79	86.49	53.73	88.29	58.26	29
	89.46		87.01		88.23		30
	87.86		88.05		87.96		31
27 folds	51.76		61.56		56.66		25
	53.67		57.92		55.80		26
	52.08		59.22		55.65		26

TABLE IV
GENERALIZATION TEST

Gp.	Cross Valid.(%)		Test (%)	Indp. Test(%)		Cd.
	MOF	Orig	MOF	MOF	Orig	
C	32.27	32.70	56.22	44.50	44.90	12
	32.59		51.89	43.00		16
	29.71		49.73	43.50		15
S	32.91	34.60	45.95	38.50	35.60	8
	33.55		42.16	38.00		7
	34.19		40.00	38.50		8
H	21.41	19.80	47.03	36.50	36.50	12
	24.92		42.70	36.50		11
	22.36		35.68	37.50		9
P	19.81	18.70	39.46	35.50	32.90	10
	21.72		37.30	34.00		11
	20.13		35.68	35.00		10
V	21.41	17.20	35.68	36.00	35.00	13
	19.17		35.68	36.50		11
	18.53		36.22	35.50		12
Z	19.81	14.60	28.65	34.00	32.90	10
	17.89		37.84	35.00		12
	20.45		36.22	31.50		10
CSH	49.84	N/A	65.95	53.50	53.90	64
PVZ	51.76		63.24	50.50		61
	51.76		61.08	55.00		50

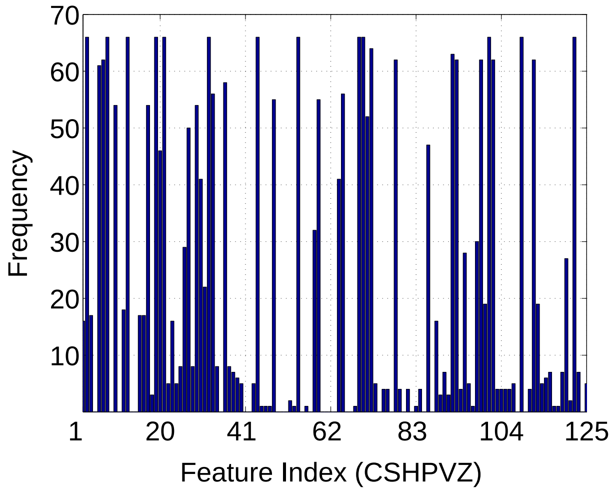

Fig. 2. Feature frequency of the solution set

VI. CONCLUSION

In this paper, we formulated the multi-class protein fold recognition problem as three objective (Cross validation accuracy on training set, Classification accuracy on testing set and Cardinality of feature subsets) optimization problem and solved it by using the multi-objective evolutionary algorithm (NSGA-II). We also extended the wrapper method by incorporating an additional testing accuracy feedback loop. This loop can spread the Pareto-optimal front with respect to the testing data set so as to provide more information. The most salient feature of the Multi-Objective Feature Analysis and Selection Algorithm is providing abundant $non - dominated$ solutions for further feature analysis and high level interpretation. We can use MOFASA to analyze the importance of features by counting their frequency among the Pareto front solutions. We can compare the bias of training and testing data sets, and find out which features or feature groups are the most sensitive to the bias. Further, the MOFASA yields higher accuracy on both cross validation and test data sets.

We used the data set in [10] in our study. The training set has less than 35% sequence similarity with the testing set. Our Q-percent prediction accuracy for 27 fold protein recognition is around 53% on cross validation data set and 60% on the test set. In independent test, six out of seven groups of features achieve high generalization performance. We examined the bias between training set and testing set by sorting the *non-dominated* solutions according to the cross validation accuracy and testing accuracy. Then we compared the results and found that features the predicted secondary structure group may cause the classifier's bias. This is our preliminary work in protein fold analysis. In the future, we will consider a more complicated situation, such as 600 fold recognition. We will also study an ensemble of different classifiers to improve the prediction accuracy.

ACKNOWLEDGMENT

The authors would like to thank Chris. H. Q. Ding at Lawrence Berkeley National Laboratory for sharing protein fold data set [9] and Chih-Jen Lin at National Taiwan University for sharing the LibSVM codes. Stanley Y. M. Shi performed this study while he was visiting Kanpur Genetic Algorithms Laboratory (KanGAL) [1] at Indian Institute of Technology Kanpur as an overseas research student in 2004.

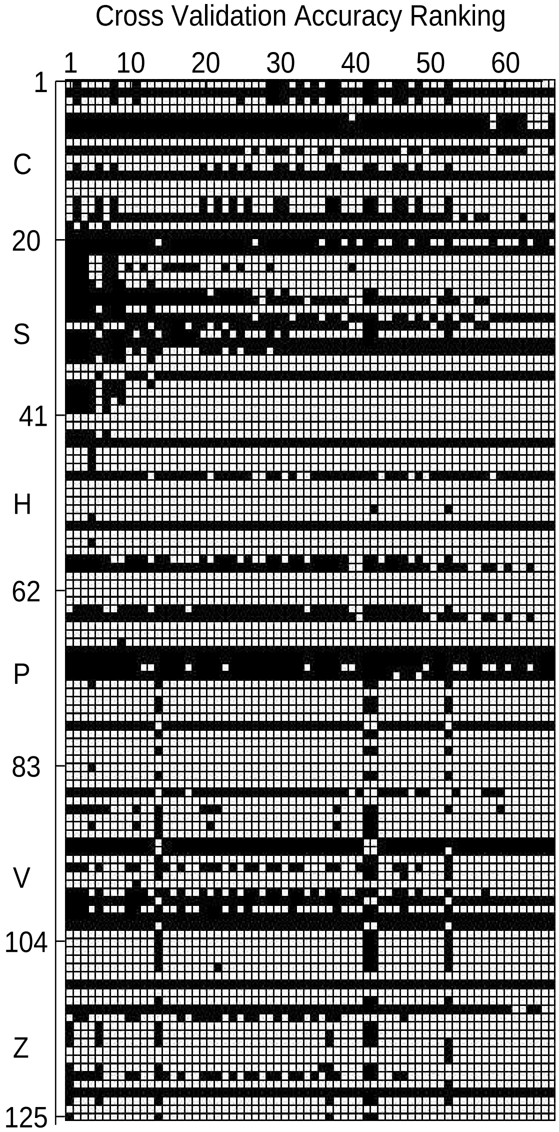

Fig. 3. *Non-dominated* Solutions sorted by cross validation results. x-axis is the solution rank and y-axis is the feature list.

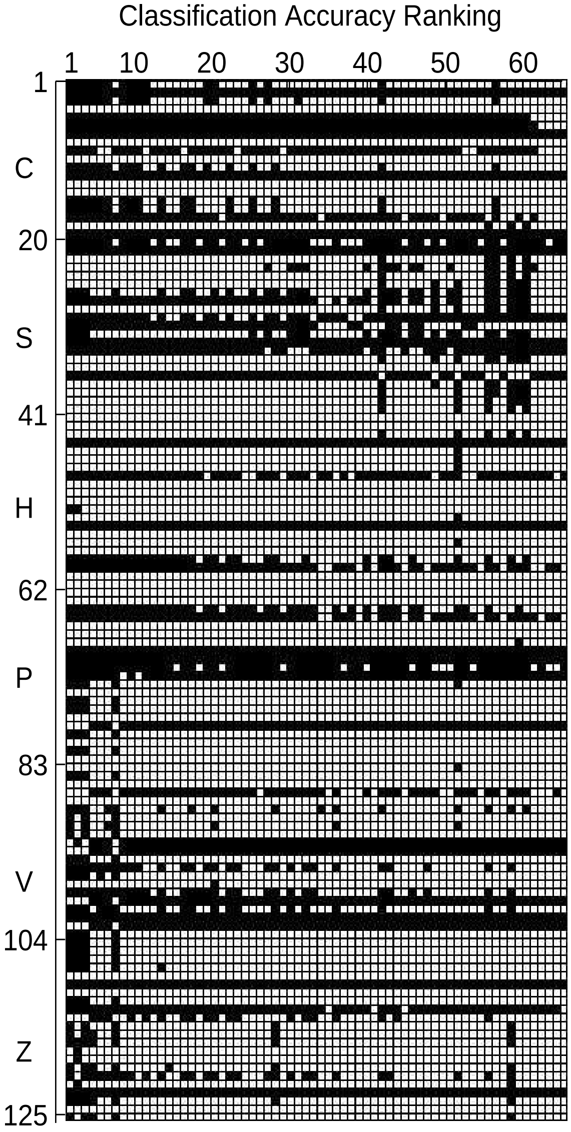

Fig. 4. *Non-dominated* Solutions sorted by classification results. x-axis is the solution rank and y-axis is the feature list.

REFERENCES

[1] Kanpur Genetic Algorithm Laboratory. $http : //www.iitk.ac.in/kangal$.

[2] A. C. Tan, D. Gilbert and Y Deville. Multi-class protein fold classification using a new ensemble machine learning approach. *Genome Informatics*, 14:206–217, 2003.

[3] P. Baldi and S. Brunak. *Bioinformatics: The Machine Learning Approach*. MIT Press, second edition edition, 2003.

[4] A. Blum and P. Langley. Selection of relevant features and examples in machine learning. *Artificial Intelligence*, 97(1-2):245–271, 1997.

[5] C. A. Orengo, A. D. Michie, S. Jones, D. T. Jones, M. B. Swindells and J. M. Thornton. CATH - a hierarchic classification of protein domain structures. *Structure*, 5:1093–1108, 1997.

[6] C. H. Wu, L. L. Yeh, H. Huang, L. Arminski, J. Castro-Alvear, Y. Chen, Z. Hu, Robert S. Ledley, P. Kourtesis, B. E. Suzek, C. R. Vinayaka, J. Zhang and W. C. Barker. The protein information resource. *Nucleic Acids Research*, 31:345–347, 2003.

[7] K. Deb. *Multi-Objective Optimization using Evolutionary Algorithms*. Wiley, Chichester, 2001.

[8] K. Deb and A. R. Reddy. Classification of two-class cancer data reliably using evolutionary algorithms. *BioSystems*, 72(1-2):111–129, 2003.

[9] C. H. Q. Ding. Protein Datasets used in the paper ”Multi-class Protein Fold Recognition Using Support Vector Machines and Neural Networks” by Chris H.Q. Ding and Inna Dubchak. $http : //crd.lbl.gov/\ cding/protein/$, 2003.

[10] C. H. Q. Ding and I. Dubchak. Multi-class protein fold recognition using support vector machines and neural networks. *Bioinformatics*, 17(4):349–58, April 2001.

[11] D. E. Goldberg. *Genetic Algorithms in Search, Optimization, and Machine Learning*. Addison-Wesley Publishing Company, Reading, MA, 1989.

[12] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov and P. E. Bourne. The protein data bank. *Nucleic Acids Research*, 28:235–242, 2000.

[13] L. Holm and C. Sander. Protein folds and families:sequence and structure alignments. *Nucleic Acids Research*, 27(1):244–247, 1999.

[14] C. W. Hsu and C. J. Lin. A comparison of methods for multi-class support vector machines. *IEEE Transactions on Neural Networks*, 13:415–425, 2002.

[15] N. Indurkhya and S. M. Weiss. Estimating performance gains for voted decision tree. *Intelligent Data Analysis*, 2(1-4):303–310, 1998.

[16] A. Jain and D. Zongker. Feature selection: evaluation, application, and small sample performance. *IEEE Transaction on Pattern Analysis and*

Machine Intelligence, 19:153–158, 1997.

[17] D. T. Jones. Genthreader: an efficient and reliable protein fold recognition method for genomic sequences. *Journal of Molecular Biology*, 287(4):797–815, April 1999.

[18] K. Deb, A. Pratap, S. Agarwal and T. Meyarivan. A fast and elitist multiobjective genetic algorithm: Nsga-ii. *Evolutionary Computation, IEEE Transactions on*, 6(2):182–197, April 2002.

[19] R. Kohavi and G. H. John. Wrappers for feature subset selection. *Artificial Intelligence*, 97(1-2):273–324, 1997.

[20] M. Kudo and J. Sklansky. Comparison of algorithms that selects features for pattern classifiers. *Pattern Recognition*, 33:25–41, 2000.

[21] C.-J. Lin. LIBSVM – A Library for Support Vector Machines. $http : //www.csie.ntu.edu.tw/\ cjlin/libsvm/index.html$, 2004.

[22] B. Rost and C. Sander. Prediction of protein secondary structure at better than 70% accuracy. *Journal of Molecular Biology*, 232:584–599, 1993.

[23] S. Y. M. Shi and P. N. Suganthan. Feature analysis and classification of protein secondary structure data. In *Internaltional Conference on Artificial Neural Networks*, volume 2714, pages 1151–1158, Istanbul, Turkey, June 2003.

[24] V. Vapnik. *Statistical Learning Theory*. Wiley, NY, 1998.

[25] Y. Cai, P. Ricardo, C. Jen and K. Chou. Application of SVM to predict membrane protein types. *Journal of Theoretical Biology*, 226(4):373–376, February 2004.