Support Vector Machine Classification of Protein Sequences to Functional Families based on Motif Selection

Danai Georgara[1,1], Katia Kermanidis[1], and Ioannis Mariolis[2]

[1] Department of Informatics, Ionian University,
7 Tsirigoti Square, 49100 Corfu, Greece
dgeorgara@yahoo.com, kerman@ionio.gr
[2] Information Technologies Institute, CERTH, Thessaloniki, Greece
ymariolis@iti.gr

Abstract. In this study protein sequences are assigned to functional families using machine learning techniques. The assignment is based on support vector machine classification of binary feature vectors denoting the presence or absence in the protein of highly conserved sequences of amino-acids called motifs. Since the input vectors of the classifier consist of a great number of motifs, feature selection algorithms are applied in order to select the most discriminative ones. Three selection algorithms, embedded within the support vector machine architecture, were considered. The embedded algorithms apart from presenting computational efficiency allowed for ranking the selected features. The experimental evaluation demonstrated the usefulness of the aforementioned approach, whereas the individual ranking for the three selection algorithms presented significant agreement.

Keywords: PROSITE database, protein classification, feature selection, machine learning

1 Introduction

Assigning putative functions to protein sequences constitutes one of the most challenging tasks in functional genomics. Protein function is often correlated with highly conserved sequences of amino-acids called motifs. Hence, motif composition is often used to assign functional families to novel protein sequences. However, many proteins usually contain more than one motif and several motifs can belong to proteins assigned to different families. Therefore, in order to reliably assign a protein to a certain family it is often required to identify motif combinations that are present in that protein. Data mining or machine learning algorithms offer some of the most effective approaches to the discovery of such unknown relationships between collections of motifs and families.

[1] Corresponding author

Wang et al. use the decision tree method for assigning protein sequences to functional families based on their motif composition [1]. The datasets used in the experiments were extracted from the PROSITE database [2]. The experimental results showed that the obtained decision tree classifiers presented a good performance.

Hatzidamianos et al. present a preprocessing software tool, called GenMiner [3], which is capable of processing three important protein databases and transforming data into a suitable format for the Weka data mining suite and MS SQL Analysis Manager 2000. A decision tree technique was used for mining protein data and the experimental results confirmed the system's capability of efficiently discovering properties of novel proteins.

Psomopoulos et al. propose a finite state automata data mining approach, which is used to induce protein classification rules [4]. The form of the extracted rules is $X \rightarrow Y$, where X is a set of motifs and Y a set of protein families. Results outperformed those obtained in [1] and [3].

Merschmann and Plastino propose a new data mining technique for protein classification based on Bayes' theorem, called highest subset probability (HiSP) [5]. To evaluate their proposal, same datasets as in [4] were used. The results have shown that the proposed method outperforms previous methods based on decision trees [3] and finite state automata [4].

Diplaris et al. present a comparative evaluation of several machine learning algorithms for the motif-based protein classification problem [6]. The results showed that a Support Vector Machine classifier provided the least mean error rate.

In the present study a Support Vector Machine classifier (SVM) is trained using a set of proteins with known function. Each protein in this set is represented by a binary input vector produced using a motif vocabulary. The classifier aims to assign novel protein sequences to one of the protein families that appear in the training set. Since the input vectors consist of a great number of motifs, Feature Selection Algorithms (FSAs) are applied in order to select the most discriminative motifs. Three FSAs, embedded within the SVM architecture, were considered. The first FSA, called Recursive Feature Elimination (RFE) [7], conducts feature selection in a sequential backward elimination manner using as a criterion the amplitude of the weights of the SVM. The second FSA is called discriminative function pruning analysis (DFPA) feature subset selection method [8]. The basic idea of the DFPA method is to learn the SVM discriminative function from training data using all input variables available first, and then to select feature subset through pruning analysis. The third, called prediction-risk-based feature selection (SBS) [9], evaluates the features by computing the change of training error when the features are replaced by their mean values. Moreover, Stepwise Discriminant Analysis (SDA) has been applied to the complete feature set, in order to compare the embedded techniques to a filter FSA.

This paper is organized as follows. In Section 2 the aforementioned feature selection methods are presented, whereas in Section 3 the experimental results are demonstrated. The paper concludes in Section 4.

2 Materials and Methods

2.1 Support Vector Machine Classification

Support Vector Machines (SVMs) [10]–[12] is a very popular choice for performing classification tasks. Since the structural risk minimization principle of SVMs chooses discriminative functions that have the minimal risk bound, SVMs are less likely to overtrain data than other classification algorithms. Because of their useful properties they were selected in this study for assigning functional families to protein sequences.

SVM is a linear machine performing binary classification. It is based on the large margin classification principle, according to which the discriminating hyperplane maximizes the margin between certain training data points of each class, called support vectors.

In some cases, using nonlinear SVMs can improve classification results. The key idea of nonlinear SVMs is mapping patterns non linearly from the input space to a transformed space, usually of higher dimensions and then perform classification in the transformed space using linear support vector machines. However, the nonlinear mapping is not explicitly performed; instead kernel functions are employed to compute the inner products between support vectors and the pattern vectors in the transformed space. The most popular kernels are Gaussian and Polynomial.

As mentioned above SVMs perform binary classification. In order to apply SVMs to multi-class problems a modification is required. In this study, the One Versus All (OVA) multi-class extension has been employed. This approach performs K binary classification between the instances of each class and the instances of all the remaining classes, where K denotes the number of different classes. OVA-SVM was preferred over other multi-class extensions not only because of its simplicity, but more importantly because it also allows for a straightforward extension of the aforementioned embedded feature selection methods.

2.2 Feature Selection

Despite the good generalization ability of SVMs, it is a good practice to reduce the feature space removing any redundant or noisy features. However, SVM feature selection based on wrapper methods [13] is inefficient. This is because it involves training a large number of SVM classifiers, with each training being computational expensive. In case of multi-class SVM the computational cost increases by a factor of K, where K is the number of different classes. On the other hand filter methods [14] that present low computational cost, do not take into account the applied classification scheme and are not very effective. A good compromise between efficiency and performance are the embedded techniques, which exploit the architecture of the classifier in order to derive the most important features. In this study, three embedded feature selection algorithms are considered.

Recursive Feature Elimination. Perhaps the most popular feature selection technique embedded to the SVMs is Recursive Feature Elimination (RFE) presented in [7]. It is based on the amplitude of the separating hyperplane's weights. In each step only the features with the highest weights are selected and the SVM classifier is retrained. In this study, the OVA multi-class extension proposed in [15] has been adopted. According to this approach in order to select the features the sum of the squared weights is calculated over the K binary classifications as shown in the following equation.

$$J_j = \frac{1}{K}\sum_{k=1}^{K}\left(w_j^k\right)^2 \quad . \tag{1}$$

In (1) J_j denotes the cost for not selecting feature j, w_j^k denotes the separating hyperplane's weight that corresponds to the j^{th} feature and the binary classifier for the k^{th} class, whereas K denotes the number of different classes.

Discriminative Function Pruning Analysis. The basic idea of the DFPA algorithm [8] is to learn the SVM discriminative function from training data using all input variables available first, and then perform pruning analysis in order to select feature subset. The pruning is implemented using a forward or backward selection procedure, combined with a linear least square estimation algorithm. The method takes advantage of the linear-in-the-parameter structure of the SVM discriminative function. In this study the backward selection procedure was preferred, since RFE also performs backward elimination of the features. Moreover, like in the RFE case, the method has been extended to apply to OVA SVM in a similar manner, i.e. averaging the selection criterion over the classes.

Prediction Risk Based Feature Selection Method. This method, originally proposed by Moody et al. [16], evaluates the features by computing the change of training error when the features are replaced by their mean values. As argued by Li et al. [9], it may be more attractive than the two previous methods, since it uses the multi-class classification SVMs directly, instead of averaging the results of the binary classifications. In that study the selection procedure was called Sequential Backward Search (SBS), and this name is also adopted for the rest of this paper.

3 Experimental Results

The dataset used for SVM training and evaluation, called genbase28, was extracted from the PROSITE database using GenMiner [3]. It contains 2934 proteins belonging to 28 classes and was also used in [4][5][6] for protein classification based on machine learning techniques. Every instance of this dataset corresponds to a certain protein and consists of the protein name, the subset of the 1185 database motifs that correspond to that protein and the name of the functional family assigned to that protein.

As discussed in [5], it is an extremely class-imbalanced dataset. Thus, pre-processing has been conducted removing proteins of the same class having identical input vectors. Then, proteins of poorly represented classes, containing less than 10 instances, were discarded. This resulted into a refined dataset of 878 proteins represented by 268 motifs and belonging to 13 classes. A vector with binary values denoting the presence (or absence) of each one of the 268 remaining motifs is assigned to each protein. These binary vectors constitute the input vectors of the SVM classifiers.

Various SVM kernels were tested on the above dataset using stratified 10-fold cross validation. More specifically, three kernel types were considered, linear, second degree polynomial, and Gaussian. Linear SVM produced the best results presenting a classification error rate of 21.07% ±3.88%, whereas analysis of variance [17] resulted to rejecting the null hypothesis that the classification error rate is the same for all three kernels at the 95% confidence level. Therefore, for the remaining experiments only the linear kernel was employed.

Linear SVM's performance was evaluated also in the case all instances were used for training, resulting to a classification error rate of 11.06%. Since the test error is about twice the training error it is safe to assume that despite the large margin property of the SVM some overfitting takes place. A common strategy to avoid overfitting is to employ feature selection techniques in order to reduce the number of features of the classifier.

A popular feature selection technique is Stepwise Discriminant Analysis [18]. It is a filter method that is easy to implement and is also computationally efficient. Using SDA, 50 motifs are selected out of the original 268, and the classification error rate drops to 16.97%, which is a significant improvement with respect to the original results. However, the classification error rate is still high compared to the training error.

As a next step, a very popular feature selection algorithm embedded to the SVM architecture, called Recursive Feature Elimination (RFE) was considered. During the experiment, in every step of the RFE algorithm only a single motif is eliminated, whereas the algorithm terminated when only one motif remained. This approach allows for ranking the motifs with respect to the order they are eliminated. The performance of SVM-RFE was evaluated for each subset of selected features using 10-fold cross-validation. The lowest classification error rate, 14.46% ±3.05%, has been achieved in case of 52 motifs. In that case the training error was 12.1% indicating that overfitting is significantly reduced.

A different feature selection algorithm embedded to the SVM architecture, called DFPA was also considered. Motif elimination and algorithm termination is similar to the RFE case, also allowing for motif ranking. The performance of SVM-DFPA was evaluated for each subset of selected features using 10-fold cross-validation. The lowest classification error rate, 14.81% ±3.46%, has been achieved in case of 45 motifs. In that case the training error was 12.22% indicating that overfitting is also significantly reduced. These results are very close to the ones achieved by RFE, while fewer motifs were selected.

Another feature selection algorithm embedded to the SVM architecture, called SBS was also considered. Like in the case of RFE and DFPA, the SBS allows for motif ranking using the same elimination and termination rules. The performance of

SVM-SBS was evaluated for each subset of selected features using 10-fold cross-validation. The lowest classification error rate, 14.92% ±3.09%, has been achieved in case of 24 motifs. In that case the training error was 13.71% indicating that overfitting is also significantly reduced. These results are very close to the ones achieved by RFE and DFPA, while very few motifs were selected.

All three embedded feature selection algorithms presented similar results. However, these results were based only on the best subset of each algorithm. Further analysis is considered regarding the performance for all subsets and the agreement of the algorithms on the ranking of the motifs. In the diagram of Fig. 1 are illustrated the learning curves for all three methods. More specifically, the classification error rate % is plotted against the number of selected motifs of each method.

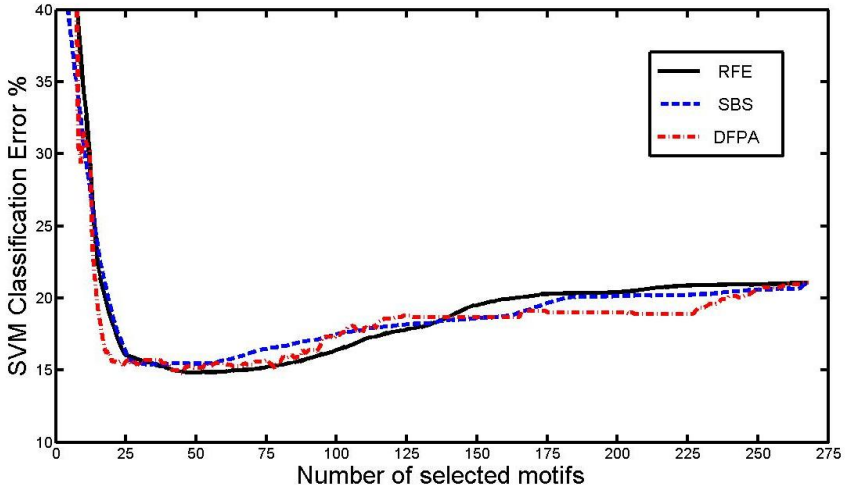

Fig. 1. Classification error rates for the three embedded FSAs, with respect to the number of selected motifs.

The three plots are very similar, when the number of motifs is not extremely low. If there are less than 12 motifs SVM-SBS outperforms the other two methods, whereas SVM-RFE presents on average slightly better results than the other two methods.

All methods present their best performance in between 20 and 80 selected motifs. Therefore, the centre of this interval, namely 50 motifs, is selected to test for the agreement of the methods on motif ranking. In that case, all methods share 37 common motifs out of the total of 50. In particular, SVM-RFE shares 46 common motifs with SVM-DFPA, whereas SVM-SBS shares 39 common motifs both with SVM-RFE and SVM-DFPA.

In Table 1 the 37 common motifs are presented, where the individual ranking of each method is also given. In the last column the median of the three individual rankings is estimated. These results indicate that there is a significant agreement

between the three methods as to which motifs of this dataset are important for protein classification.

Table 1. The 37 motifs common in all three methods' 50 motifs selections

	Motif	RFE rank	DFPA rank	SBS rank	Median rank
1	**PS00022**	3	2	2	2
2	**PS01186**	1	3	16	3
3	**PS50114**	4	13	3	4
4	**PS50109**	5	6	6	6
5	**PS50071**	6	8	1	6
6	**PS00561**	8	7	7	7
7	**PS00562**	7	15	8	8
8	**PS00193**	9	9	5	9
9	**PS00010**	2	10	17	10
10	**PS00192**	10	5	15	10
11	**PS50322**	15	4	10	10
12	**PS00188**	11	16	9	11
13	**PS00187**	12	14	12	12
14	**PS50079**	16	12	4	12
15	**PS00025**	14	17	11	14
16	**PS50099**	17	1	14	14
17	**PS00177**	13	18	19	18
18	**PS50312**	36	20	18	20
19	**PS50326**	20	32	21	21
20	**PS50830**	21	33	20	21
21	**PS50280**	22	37	23	23
22	**PS50318**	37	24	13	24
23	**PS50089**	19	26	27	26
24	**PS50129**	18	27	34	27
25	**PS01040**	25	29	37	29
26	**PS50044**	29	34	28	29
27	**PS50313**	47	30	22	30
28	**PS50324**	33	11	32	32
29	**PS50316**	45	22	33	33
30	**PS50016**	31	35	36	35
31	**PS50325**	50	23	35	35
32	**PS50215**	39	36	24	36
33	**PS00402**	26	38	40	38
34	**PS00136**	27	39	41	39
35	**PS00875**	32	40	39	39
36	**PS00012**	28	44	42	42
37	**PS50303**	49	43	30	43

It should be mentioned that the FSA methods produce lower classification error rates than those presented in [4], [5] and [6]. However, the focus of this study is on motif selection and ranking, and pre-processing was performed on that basis resulting to a smaller dataset with less than 28 classes. Therefore, it does not make much sense

to perform direct comparisons of the classification results to those of the previous studies and this is why the experimental results do not include such comparisons.

All experiments were conducted using Matlab 7.8.0 programming environment. SVM classification was performed using the SVM-KM toolbox [18], whereas the embedded feature selection methods were implemented by the authors.

4 Conclusion

In this work three FSAs embedded to the SVM architecture, were employed and evaluated on a protein classification task. The classification scheme was used to assign protein sequences to functional families, based on binary features denoting the presence or absence of motifs in their sequences.

A real dataset extracted from the PROSITE database has been employed for the evaluation of the aforementioned schemes. The experimental results demonstrated that all three feature selection methods can greatly reduce the test error of the used data set. The reduction magnitude is about 7 % of the test error on the total feature set. This indicates that the data set has some redundant or even noisy motifs, which decrease the performance of the learning machine. Moreover, the three feature selection algorithms presented significant agreement on their rankings. Therefore, prior to protein sequence classification, even with robust classifiers like SVM, it is highly recommended that the motifs comprising the feature vectors should be carefully selected. In future work the fusion of the individual ranking results will be studied in order to derive even more robust selections, improving even more the generalization ability of the employed classifier.

Acknowledgments. The authors wish to thank Dr. Luiz Merschmann for providing the experimental dataset and the anonymous reviewers for their useful comments and suggestions.

References

1. Wang, D., Wang, X., Honavar, V., Dobbs, D.: Data-driven Generation of Decision Trees for Motif-based Assignment of Protein Sequences to Functional Families. In: Atlantic Symposium on Computational Biology, Genome Information Systems & Technology (2001)
2. PROSITE, http://prosite.expasy.org/
3. Hatzidamianos, G., Diplaris, S., Athanasiadis, I., Mitkas, P.A.: GenMiner: A Data Mining Tool for Protein Analysis. In: 9th Panhellenic Conference on Informatics, Thessaloniki, Greece, pp. 346--360 (2003)
4. Psomopoulos, F., Diplaris, S., Mitkas, P.A.: A Finite State Automata Based Technique for Protein Classification Rules Induction. In: 2nd European Workshop on Data Mining and Text Mining for Bioinformatics, Pisa, Italy, pp. 54--60 (2004)
5. Merschmann, L., Plastino, A.: A Lazy Data Mining Approach for Protein Classification. IEEE Transactions on Nanobioscience 6, 36--42 (2007)

6. Diplaris, S., Tsoumakas, G., Mitkas P.A., Vlahavas, I.: Protein Classification with Multiple Algorithms. In: 10th Panhellenic Conference in Informatics, Volos, Greece, November 2005. LNCS, vol. 3746, pp. 448--456. Springer-Verlag (2005)
7. Guyon, I., Weston, J., Barnhill, S., Vapnik, V.: Gene Selection for Cancer Classification Using Support Vector Machines. Machine Learning 46, 389--422 (2002)
8. Mao K.Z.: Feature Subset Selection for Support Vector Machines through Discriminative Function Pruning Analysis. IEEE Transactions on Systems, Man, and Cybernetics 34, 60--67 (2004)
9. Li, G., Yang J., Liu, G., Li, X.: Feature Selection for Multi-class Problems Using Support Vector Machines. In: 8th Pacific Rim International Conference on Artificial Intelligence (PRICAI-04), LNCS, vol. 3157, pp. 292--300 (2004)
10. Vapnik, V.: The Nature of Statistical Learning Theory. second ed., Springer-Verlag, New York (1999)
11. Scholkopf, B.: Support Vector Learning. Oldenburg-Verlag, Munich, Germany (1997)
12. Burges, C.: A Tutorial on Support Vector Machines for Pattern Recognition, Data Mining Knowledge Discovery 2, 121--167 (1998)
13. Kohavi, R., John, G.H.: Wrappers for Feature Subset Selection. Artificial Intelligence 97, 273--324 (1997)
14. Kudo, M., Sklansky, J.: Comparison of Algorithms that Select Features for Pattern Classifiers. Pattern Recognition 33, 25--41 (2000)
15. Zhou, X., Tuck, D.P.: MSVM-RFE: Extensions of SVM RFE for Multiclass Gene Selection on DNA Microarray Data. Bioinformatics 23, 1106--1114 (2007)
16. Moody, J., Utans, J.: Principled Architecture Selection for Neural Networks: Application to Corporate Bond Rating Prediction. In: Moody, J.E., Hanson, S.J., Lippmann, R.P., (eds.) Advances in Neural Information Processing Systems 4, pp. 683--690. Morgan Kaufmann Publishers, Inc. (1992)
17. Hogg, R.V., Ledolter, J.: Engineering Statistics. MacMillan, New York (1987)
18. Einslein, K., Ralston A., Wilf H.S.: Statistical Methods for Digital Computers. John Wiley & Sons, New York (1977)
19. Canu, S., Grandvalet, Y., Guigue, V., Rakotomamonjy, A., SVM and Kernel Methods Matlab Toolbox. Perception Systèmes et Information, INSA de Rouen, Rouen, France (2005)