The Role of Eigen-matrix Translation in Classification of Biological Datasets

Hao Jiang
Advanced Modeling and Applied Computing Laboratory
Department of Mathematics
The University of Hong Kong, Hong Kong
Email: haohao@hkusuc.hku.hk

Wai-Ki Ching
Advanced Modeling and Applied Computing Laboratory
Department of Mathematics
The University of Hong Kong, Hong Kong
Email: wching@hku.hk

***Abstract*—Driven by the challenge of integrating large amount of experimental data obtained from biological research, computational biology and bioinformatics are growing rapidly. Machine learning methods, especially kernel methods with Support Vector Machines (SVMs) are very popular tools. In the perspective of kernel matrix, a technique namely Eigen-matrix translation has been introduced for protein data classification. The Eigen-matrix translation strategy owns a lot of nice properties while the nature of which needs further exploration. We propose that its importance lies in the dimension reduction of predictor attributes within the data set. This can therefore serve as a novel perspective for future research in dimension reduction problems.**

***Index Terms*—Classification; Dimension Reduction; Eigen-matrix translation; Kernel Method (KM); Support Vector Machine (SVM).**

I. Introduction

Protein function prediction can be viewed as a classification problem from the point of view of computer science [1]. With the increasing popularity of kernel-based methods for pattern classification [2], [3], a quantity of string kernels have been proposed. For example, Spectrum Kernel [4], MisMatch Spectrum Kernel [5] and kernel Based on Latent Semantic Analysis [6]. The Weighted Degree Kernel [7] has been applied in the recognition of alternatively spliced exons which rewards with a score on the length of the matching substrings. However, the above string kernels do not admit similarity among different features and this may lead to a biased result from the physico-chemical perspective. To this end, recently AAindex Based Kernel has been developed [8] for pairwise protein homology detection. A novel kernel based on the K-Spectrum Kernel, which incorporates both physico-chemical and biological information in the protein sequences for the captured protein classification problem has been proposed [9]. The main innovation of the proposed method lies in the Eigen-matrix translation technique. The main idea and effect of the technique is to add a rank one Symmetric Positive Semi-Definite (SPSD) matrix to the original kernel matrix. In general, the effect of the technique is to change one of the zero eigenvalues to a positive one. But at the same time it does not bring much perturbation to the original positive eigenvalues that are critical to fulfilling the classification problem, see for instance [10, Weyl's theorem].

We shall investigate the major role of the Eigen-matrix translation. Focusing on the characteristics of the data sets, the number of data instances is 40 to 50 while the number of features can be up to 7000 to 8000. This gives us a clue on the potential power of the Eigen-matrix translation in dimension reduction. High-dimensional data sets give us many mathematical challenges in computation as well as opportunities for new theoretical developments [11]. Traditional algorithms in machine learning and pattern recognition applications are often susceptible to the well-known problem of the curse of dimensionality [12], while certain computationally expensive novel methods [13] can construct predictive models with high accuracy from high-dimensional data. Therefore it is still of interest in many applications to reduce the dimension of the original data set prior to any analysis.

In this paper, we put forward a novel perspective of feature selection technique. We suggest that the major role of Eigen-matrix translation lies in the feature selection within the huge size of feature dimensions. The remainder of the paper is structured as follows. In Section 2, we give the formulation of the captured problem. The relationship of the Eigen-matrix translation and feature selection is then illustrated. Computational results on real biological data sets are then given to verify the correctness of the proposed framework and models in Sections 3 and 4. Finally, concluding remarks are given in the last section.

II. Methodology

In this section, we present the relationship between the Eigen-matrix translation technique and feature selection in binary classification. Since the Eigen-matrix translation technique was first established on the spectrum kernel, it is more coherent to give a brief introduction on spectrum kernel. Details will be left for the illustration of internal connections of Eigen-matrix translation and feature selection.

A. Spectrum Kernel

In the protein classification, spectrum kernel is one of the outstanding representatives in string kernels. In K-spectrum kernel, the input space χ consists of all finite K-length sequences from the alphabet $\Re$ and $|\Re| = n$. We assume the input data set contains N sequences $\{p_1, p_2, \ldots, p_N\}$. The

978-1-4673-2560-8/12/$31.00 ©2012 IEEE

K-spectrum of an input sequence p_i is the set of all the K-length (contiguous) subsequences that it contains. The K-mer representation of the sequence is denoted through a feature map from χ to R^{n^K}. Here

$$x_i^K = [x_{1i}^K, x_{2i}^K, \ldots, x_{n_K i}^K]^T$$

and x_{li}^K is the occurrence of lth K-mer in the input data p_i, n_K is the dimensionality of features. If V_K is the K-mer representation matrix for the whole input data set of dimensionality $n_K \times N$, then its K-spectrum kernel can be expressed as follows: $Ker_K = V_K^T \cdot V_K$.

B. Eigen-matrix translation

The Positive Semi-Definite (PSD) property is guaranteed by

$$Ker_K = V_K^T \cdot V_K.$$

The Eigen-matrix translation technique is an effective procedure to improve classification accuracy. It involves two steps:

1) The Eigenvalue Decomposition:

$$Ker_K = X \cdot P \cdot X^T$$

where X is the orthogonal matrix containing all the column eigenvectors of the matrix Ker_K and P is the diagonal matrix containing all the corresponding eigenvalues of Ker_K, see for instance [10].

2) The Eigen-matrix translation technique:

$$Ker_{new} := X \cdot [P + \lambda[1, 1, \ldots, 1]^T \cdot [1, 1, \ldots, 1]] \cdot X^T.$$

C. Problem Formulation

The Eigen-matrix translation technique [14] has shown to be successful in improving classification accuracy in protein classification problems where the problems have few data instances but huge size of features. Here we assume the number of data instances is n, the number of features for each data instance is p, then $p \gg n$. Let S be a $p \times p$ matrix, then the original kernel matrix before making the Eigen-matrix translation can be re-written in the following form (from the perspective of similarities among different features):

$$Ker_K = V_K^T \cdot S \cdot V_K.$$

Here S is the identity matrix of dimension $p \times p$. In other words, the spectrum kernel assumes no similarity between two different features. Each feature is regarded equally important and all of the features are used for classification.

We note that

$$Ker_{new} - Ker_K = X \cdot \lambda[1, 1, \ldots, 1]^T \cdot [1, 1, \ldots, 1] \cdot X^T.$$

If we write $Ker_{new} - Ker_K$ in the following form:

$$Ker_{new} - Ker_K = V_K^T \cdot \Delta S \cdot V_K$$

where ΔS is also a diagonal matrix of size $p \times p$. The nonzero entries among the p diagonal entries are important in classification. On the one hand, for the zero entries, it means the corresponding features are invariable to the Eigen-matrix translation technique. On the other hand, for those nonzero ones, they indicate either strengthening or impairing the effect of features in classification. In this context, we have

$$Ker_{new} = V_K^T \cdot [S + \Delta S] \cdot V_K.$$

As a matter of fact, the above assumptions can be realized with the following construction procedures. Since $Ker_{new} - Ker_K$ is a rank one matrix, it can be represented as follows:

$$Ker_{new} - Ker_K = Y \cdot Y^T$$

where $Y = [y_1, y_2, \ldots, y_n]^T$ and $y_i = \sqrt{\lambda} \sum_{j=1}^{N} x_{ij}$, $i = 1, 2, \ldots, n$. If we denote $\Delta S = \mathrm{diag}(S_{11}, \ldots, S_{pp})$, the problem can then be transformed to the problem of solving a linear system of $n(n+1)/2$ equations and p unknowns:

$$\begin{bmatrix} v_{11}^2 & \cdots & v_{1p}^2 \\ \vdots & \vdots & \vdots \\ v_{n1}^2 & \cdots & v_{np}^2 \\ v_{21}v_{11} & \cdots & v_{2p}v_{1p} \\ \vdots & \vdots & \vdots \\ v_{n1}v_{11} & \cdots & v_{np}v_{1p} \\ \vdots & \vdots & \vdots \\ v_{n1}v_{n-11} & \cdots & v_{np}v_{n-1p} \end{bmatrix} \begin{bmatrix} S_{11} \\ S_{22} \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ S_{pp} \end{bmatrix} = \begin{bmatrix} y_1^2 \\ \vdots \\ y_n^2 \\ y_2 y_1 \\ \vdots \\ y_n y_1 \\ \vdots \\ y_n y_{n-1} \end{bmatrix}.$$

In order to obtain a unique solution, we utilize the linear system solver under the condition of least square errors. Suppose

$$I_{irrelevant} = \{i, S_{ii} = 0, i \in \{1, 2, \ldots, p\}\}$$

and

$$I_{relevant} = \{i, S_{ii} \neq 0, i \in \{1, 2, \ldots, p\}\}.$$

Hence we have

$$I_{irrelevant} \cap I_{relevant} = \emptyset, I_{relevant} \cup I_{relevant} = \{1, 2, \ldots, p\}.$$

The indices in $I_{relevant}$ represent the features essential for classification. Hence we choose those features for classification. We note that the new kernel in this context becomes

$$Ker_1 = [V_K(I_{relevant}, :)]^T \cdot V_K(I_{relevant}, :).$$

III. Experiment Results

A. Data Source

Since the Eigen-matrix translation technique was first applied to protein classification, we employ the same three sets of glycan-binding related protein data to demonstrate the effectiveness of our proposed scheme. Glycan structures, lectin-glycan binding affinity, lectin sequences are retrieved from the the glycan database of Functional Glycomics Gateway (CFG) [15]. We assume a lectin binds to a glycan if the binding affinity exceeds 10000. Here we focus on the glycan structures with a relatively large number (≥ 20) of binding lectins and obtained three qualified glycans. The glycan structures are illustrated in Table I.

In the captured three glycan structures, glycan-binding protein prediction can be regarded as a classification problem to assess the binding property of a protein sequence. In Glycan 1

TABLE I: 3 Glycan Structures

Glycan 1	[3OSO3]Galb1-3GalNAca-Sp8
Glycan 2	NeuAca2-3(NeuAca2-3(GalNAcb1-4)Galb1-4Glcb-Sp0)
Glycan 3	NeuAca2-8NeuAca2-8NeuAca2-8NeuAca2-3(GalNAcb1-4)Galb1-4Glcb-Sp0

related data set, we have 23 positive data. In Glycan 2, data set contains 22 positive data. Similarly, 20 positive data constitute the positive part of the third data set. The same number of the negative data for each data set was then chosen to ensure the balance of positive and negative data. The dimensions of the features of the three data sets are huge, and they are 8202, 7377 and 7815 respectively. These are typical examples of small n and large p.

B. Experiments

The focus of our study is the role of the Eigen-matrix translation technique in classification, we therefore concentrate on the data sets whose classification accuracies are improved after making an Eigen-matrix translation. In the following experiments, 4-spectrum kernel is adopted for making comparisons. The reason for employing 4-mer as a feature is following. It was suggested by a prior research in [4], [5] and [16] that 4-mers is superior for string kernel.

In the first step, we select the qualified data sets whose classification accuracies improve merely after using the Eigen-matrix translation technique. After comparing the performance with the 4-spectrum kernel, two data sets are then selected, they are Glycan 2 and Glycan 3 related protein data sets. Results are illustrated in the figures in [17]. For the purpose of illustration, we use one data set: Glycan 3 Related Data set. In performing the Eigen-matrix translation, $\lambda = 0.1$ was chosen as previous sensitivity analysis in [9] has guaranteed that for λ in [0.01,1], better classification result can be obtained. To some extent it renders λ a free tuning parameter. We therefore employ 0.1 as the value of λ.

However, different value of λ may result in different effects on selection of predictor variables in the final model. Therefore, it is necessary to conduct a sensitivity analysis on λ for feature selections. We use the same step size as in [9], in addition, we use 0.001 as stepsize for λ in $[0.001, 0.01]$. For λ in [0.001,1], the same number of features are selected (data is not shown). Hence we observe the variation of λ would not change the final selected features. Thus in the following settings, we will use $\lambda = 0.1$. Secondly, in order to demonstrate the validity of the proposal, numerical experiments are implemented on the selected data sets. The effectiveness of our method is then evaluated through comparison with the 4-spectrum kernel method in terms of AUC values. The experimental results indicate that the proposed method is an effective tool for dimension reduction in classification. For Glycan 3 related data set, 110 out of 7815 features are selected for classification. The AUC values are also larger than the original 4-spectrum kernel, see Figure 1 for instance. This further confirms the effectiveness of our developed scheme for classification.

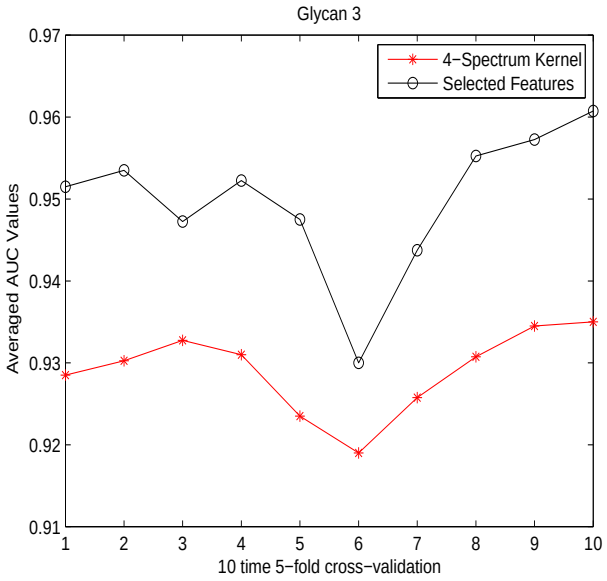

Fig. 1: Comparison of performance on Glycan 3 Related Data in mer4. This figure compares the performance of 4-spectrum kernel with the new kernel constructed on selected features where the features are selected through the framework of Eigen-matrix translation Technique, with our proposed method. 10 times 5-fold cross validation were performed as well, introducing Averaged AUC values for criterion. For Glycan 3 Related Data set, with our proposed method, the new kernel on the selected 110 features out of 7815 features achieves better classification accuracy. The averaged AUC values for 4-spectrum kernel is about 0.93 while for the kernel on selected features, averaged AUC values improve to 0.95.

We also concerned that if the Eigen-matrix translation technique, as a feature selector, can outperform other feature selectors. Therefore, we conduct study on the comparison of our model with other feature selectors. We select two famous feature selectors for comparison. One ranks the features in the data set using an Two-sample T-test for binary classification. The other criterion tries to maximize the area under the ROC curve to assess the significance of every feature for separating two labeled groups. Since these two feature ranking algorithms cannot determine the number of significant features automatically, we use the same number of features in our proposed model for comparison study. Figure 2 clearly illustrates the superiority of our proposed model as a feature selector.

C. Simulation Study

In this subsection, we conduct a simulation study to demonstrate that the Eigen-matrix translation technique is indeed a good feature selector. We generate the dataset with 40 data instances and 2000 features, 20 of them are positive and 20 are negative. Among the 2000 features, 1970 features are the same for all the data instances. For simplicity, we use 0 for all feature values. Then we perturb the 30 features with randomly generated vectors of length 10. Actually this kind of dataset has 30 important features for classification. All the other 1970 features can be regarded as noises as they bring no difference between positive and negative data. After generating the dataset, we apply the Eigen-matrix translation technique

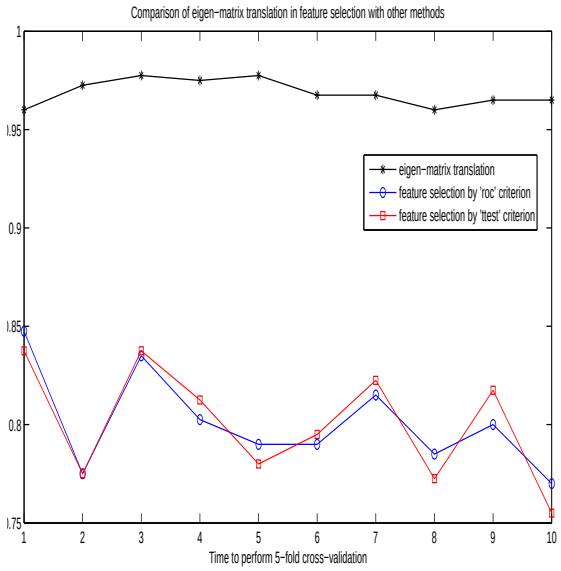

Fig. 2: Comparison of different feature selectors on Glycan 3 Related Data. The solid line with '*' is the performance of selected features by eigen-matrix translation based model , with solid line in '○' standing for feature selectors by 'roc' criterion. Solid line in '□' stands for feature selectors by 'ttest' criterion. Our model performs best in the 10 times 5-fold cross-validations. Around 0.96 AUC values were achieved in average. For feature selectors by 'roc' and 'ttest' criterion, similar performance can be detected, reaching on average 0.8 AUC values.

to the original linear kernel constructed and select a subset of features for classification. The results indicate that that Eigen-matrix translation can indeed select the exact 30 features for classification.

IV. Discussions

Experiments have further validated the effective role of our proposed scheme in feature selection. The selected features used in classification show their superiority when compared to the whole set of features. In Glycan 3 Related data set with 4-mer features, the number of features have reduced from 7815 to 110. This is a drastic improvement particularly in data sets with massive number of features. The Eigen-matrix translation played an important role in the feature selection, the newly constructed kernel after feature selection outperformed the original k-mer kernel in terms of classification accuracy. As a conclusion, through mathematical formulation, we have constructed relationship between Eigen-matrix translation and feature selection. And the effectiveness of the model is assured by the experiment results.

V. Concluding Remarks

In this paper we investigate the major role of the Eigen-matrix translation technique in SVM for classification. Through the mathematical formulation of the problem, we suggest the effective role of the technique in feature selection. Different from the traditional algorithms in feature selection, our method realizes the task of dimension reduction in an automatic way. And the sparsity of the selected subset of features is guaranteed. Experimental results further validate the effectiveness of our proposed method. This also provides another perspective in dimension reduction for data sets of huge size of features.

Acknowledgment

Research supported in part by HKU Strategy Research Theme fund on Computational Sciences, National Natural Science Foundation of China Grant No. 10971075 and Guangdong Provincial Natural Science Grant No. 9151063101000021.

References

[1] K.M. Borgwardt and H.P. Kriegel, *Kernel Methods for Protein Function Prediction*, AFP-SIG. Detroit, USA: Oxford, 2005.
[2] J. Shawe-Taylor and N. Cristianini, *Kernel Methods for Pattern Analysis*, Cambridge: Cambridge University Press, 2004.
[3] T.Jaakola, M.Diekhans and D.Haussler, "A Discriminant Framework for Detecting Remote Protein Homologies," *Journal of Computational Biology*, vol.7, pp.95-114, 2000.
[4] C. Leslie, E. Eskin, A. Cohen and W.S. Noble, "The Spectrum Kernel: A String Kernel for SVM Protein Classification," in *Proceedings of the Pacific Biocomputing Symposium*, 7, 2002, pp.566-575.
[5] C. Leslie, E. Eskin, J. Weston and W.S. Noble, "Mismatch String Kernel for Discriminative Protein Classification," *Bioinformatics*, vol.20, pp.467-476, 2004.
[6] Y.S. Yuan, L. Lin, Q.W. Dong, X.L. Wang and M.H. Li, "A Protein Classification Method Based on Latent Semantic Analysis," in *Proceedings of the 2005 IEEE Engineering in Medicine and Biology 27th Annl. Conf.* vol.7, 2005, pp. 7738-7741.
[7] G. Ratsch, S. Sonnenburg and B. Scolkopf, "RASE: Recognition of Alternatively Spliced Exons in c. elegans," *Bioinformatics*, vol.21, pp. 1369-1377, 2005.
[8] B.J.M. Webb-Robertson, K.G. Ratuiste and C.S. Oehmen, "Physicochemical Property Distributions for Accurate and Rapid Pairwise Protein Homology Detection," *BMC Bioinformatics*, vol.11, pp.145, 2010.
[9] H. Jiang and W. Ching, "Physico-chemically Weighted Kernel for SVM Protein Classification," in *Proceedings of the 2nd International Conference on Biomedical Engineering and Computer Science (ICBECS 2011)*, 23-24 April, Wuhan, China, 2011, pp.12-15.
[10] R. Horn and C. Johnson, *Matrix Analysis*, Cambridge : Cambridge University Press, 1985.
[11] D.L. Donoho, "High-dimensional Data Analysis: The Curses and Blessings of Dimensionality," in *American Mathematical Society Conference of Math Challenges of the 21st century*, Los Angeles, August 6-11, 2000.
[12] R. Bellman. *Adaptive Control Processes: A Guided Tour*, Princeton, New Jersey: Princeton University Press, 1961.
[13] L. Breiman, "Random Forests," *Machine Learning*, vol.45, pp.5-32, 2001.
[14] H.Jiang and W.K.Ching, "Kernel Techniques in Support Vector Machines for Classification of Biological Data," *International Journal of Information Technology and Computer Science*, vol.3, pp.1-8, 2011.
[15] Functional Glycomics Gateway. Available at http://www.functionalglycomics.org.
[16] Y. Yang, L. Lin, Q. Dong, X. Wang and M. Li, "Remote Protein Homology Detection Using Recurrence Quantification Analysis and Amino Acid Physicochemical Properties," *J. Theor. Biol.*, vol.252, pp. 145-154, 2008.
[17] http://hkumath.hku.hk/∼wkc/papers/ieeeadditionalfile1.pdf