Protein Secondary Structure Prediction
A Matrix Completion Approach

Priyank Patel
UIN: 00985578
Email: ppatel@cs.odu.edu

Department of Computer Science
Old Dominion University

Advisor

Yaohang Li
Email: yaohang@cs.odu.edu

Department of Computer Science
Old Dominion University

Abstract:

In this project, we present a novel predictive model for protein structure prediction based on matrix completion. Our purpose is to introduce a completely new approach to the protein structure prediction problem and evaluate the performance of matrix completion approach on protein data. For the evaluation purpose we compared the results of theses model with the results generated by protein structure prediction model based on neural network with almost 90% accuracy. We have also explained probability refinements used to estimate and normalize our predicted probability from the approximated matrix to make them comparable with the neural network model. The prediction accuracy achieved by our model is very low compared to the other available model models. Furthermore, we analyze the pitfalls that were encountered during our research so others can improve upon our methodology for future research in this field.

Introduction:

Protein secondary structure prediction remains an important step on the way to full tertiary structure prediction in computational biology. A variety of approaches have been proposed to derive the secondary structure of a protein from its amino acid sequence as a classification problem. Beginning with the seminal work of Qian and Sejnowski [1], many of these methods have utilized neural networks. A major improvement in the prediction accuracy of these methods was made by Rost and Sander [2], who proposed a prediction scheme using multi-layered neural networks, known as PHD. The key novel aspect of this work was the use of evolutionary information in the form of profiles derived from multiple sequence alignments instead of training the networks on single sequences. The key novel aspect of this work was the use of evolutionary information in the form of profiles derived from multiple sequence alignments instead of training the networks on single sequences. Another type of alignment profile, position-specific scoring matrices (PSSM) derived by the iterative search procedure PSI-BLAST [3], has been used in neural network prediction methods to achieve further improvements [4, 5].

In this approach, we tried to predict protein secondary structures from available profiles using matrix completion. Initially, to generate input data we gathered protein feature profiles from previous experiments conducted by Dr Yohang Li, for some of these proteins profiles the secondary structures were known already. We arranged this information into a big matrix form, and applied matrix completion to predict the secondary structure information of the unknown proteins. To check the performance of the matrix completion on these data, we compared the recovered values of the matrix with the results of the neural network experimental results conducted on the same data with very high prediction accuracy. For this comparison we first calculated the probability values for classes, class $\alpha-$helix $,\beta-$sheets and coils. Then we normalized these probabilities to make them comparable to the neural network results.

Matrix Completion Approach:

There is a rapidly growing interest in the recovery of an unknown low-rank or approximately low-rank matrix from very limited information. This problem occurs in many areas of engineering and applied science such as machine learning [6-8], control [9] and computer vision, see [10]. For example, consider the problem of recovering a data matrix from a sampling of its entries. In the area of recommender systems, users submit ratings on a subset of entries in a database, and the vendor provides recommendations based on the user's preferences. Because users only rate a few items, one would like to infer their preference for unrated items; this is the famous Netflix problem [11]. The objective of matrix completion is to recover the missing (unknown) entries of an incomplete matrix from a small subset of observed ones [11-13]. The matrix recovered from matrix completion is mostly low rank or nearly low rank matrix under the assumption that generally the most actions of matrices are effected by only a few factors in real-life applications.

Let M denote an incomplete matrix and Ω be a set of indices of observed entries; the matrix completion problem is then defined as finding a low-rank solution X to the following optimization problem.

$$\text{minimize } \|X\|*$$

$$\text{subject to } \mathcal{P}_{\Omega}(X) = \mathcal{P}_{\Omega}(M)$$

Where, $\|\cdot\|*$ is the nuclear norm which is the sum of singular values and $\mathcal{P}_{\Omega}$ is the projection operation defined as

$$\mathcal{P}_{\Omega}(\mathrm{X})_{ij} = \begin{cases} M_{ij} & \text{if } (i,j) \in \Omega \\ 0 & \text{if } (i,j) \notin \Omega \end{cases}$$

Many numerical algorithms have been developed in the literature to solve the above matrix completion problem. For example, convex optimization algorithms based on Semi-definite Programming to fill out the missing matrix [11, 12] and the Singular Value Thresholding (SVT) algorithm to efficiently approximate the optimal result [13].

We tried to formulate the protein secondary structure prediction problem in to the matrix completion problem by arranging the known protein structure profiles into a matrix together with the unknown ones, which is to be predicted. In Figure-1 we tried to give an idea about the matrix completion procedure. In the matrix red region indicate the combination of known profiles and the unknown protein profiles and the blue region which to be completed using matrix completion indicates unknown protein classes.

Figure-1: Matrix Completion Process

In the dataset matrix (M) we built for this project is of size 48520 x 320, meaning there were total of 48520 proteins we considered in the experiment and for each of this protein we considered a profile of 315 features in it. Next 3 columns of the matrix are indicating of protein structures like, β −sheets and α −helix respectively and the remaining 2 were other parameters. Out of the 48520 proteins, classes for 35831 protein profiles were known and classes for remaining 12689 protein profiles were to be predicted using the matrix completion.

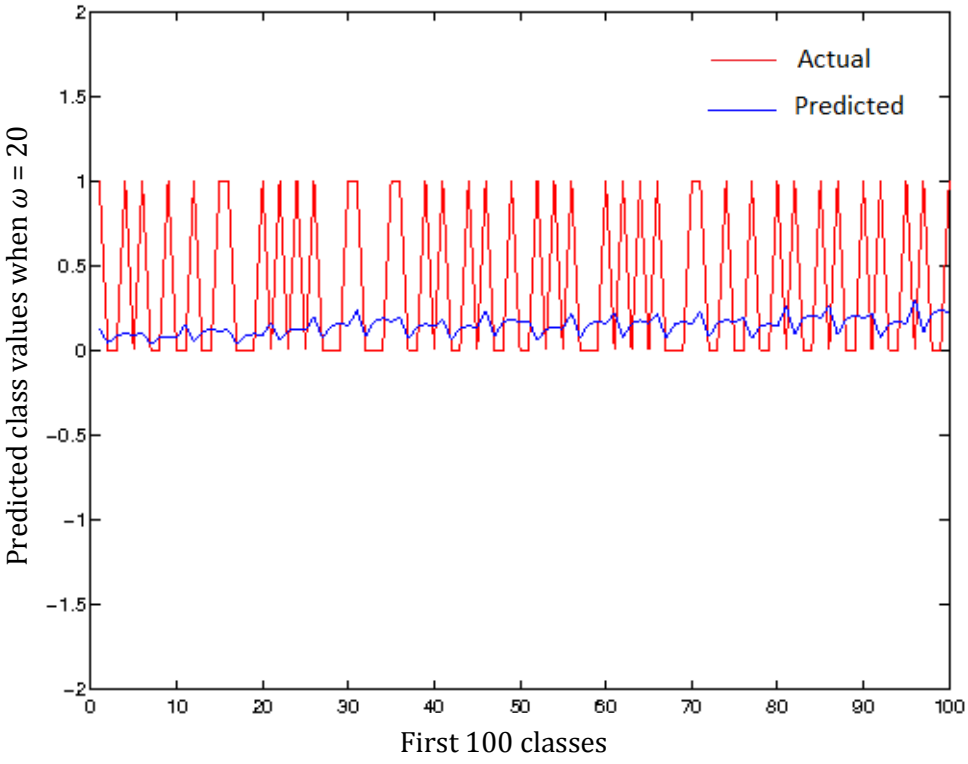
Method:

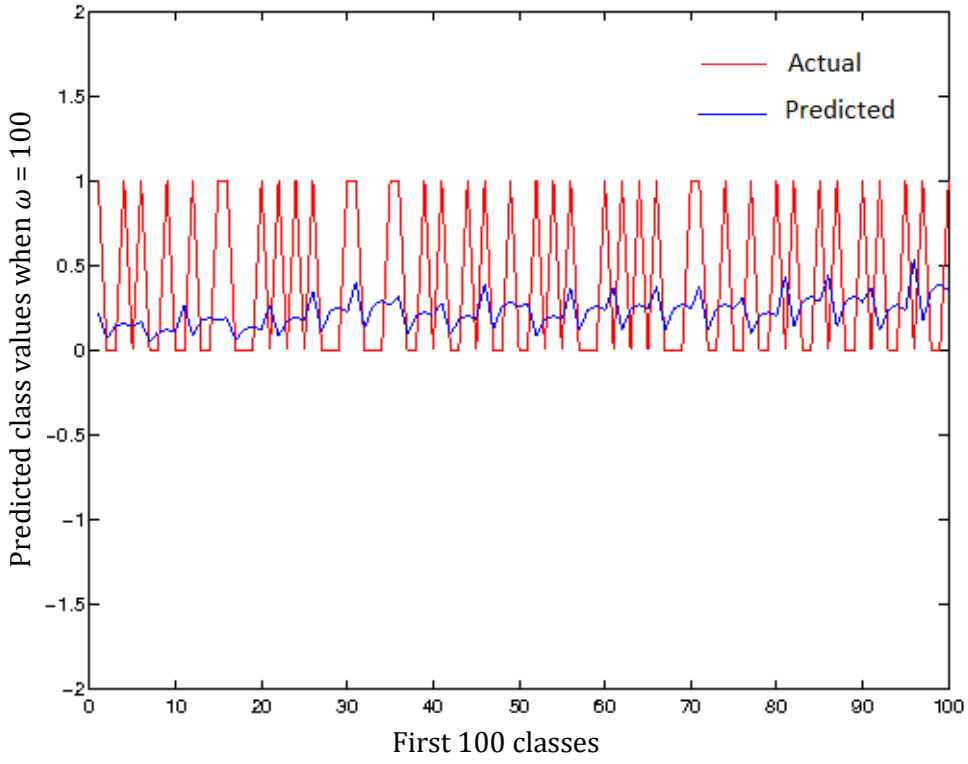
We applied SVT algorithm [13] on matrix M to complete the unknown classes in the matrix. Here, Ω is assigned with a set of indices of known values in the matrix. The SVT algorithm seeks a low-rank matrix X to minimize the following Lagrange dual function,

$$\tau\|X\|* + \frac{1}{2}\|\mathcal{P}_{\Omega}(X) - \mathcal{P}_{\Omega}(\mathrm{M})\|_F^2$$

Where, $\mathcal{P}_{\Omega}$ is the projection operation and τ is a Lagrange multiplier trading off between the nuclear and Frobenius norm. In general, suppose the matrix to recover is of size $m \times n$, the value of τ is set to be a factor of $\sqrt{mn}$, such that $\tau = \omega\sqrt{mn}$, where ω is a positive number.

A difficulty in applying the SVT algorithm to March Madness prediction is that not all values of ω can make the SVT algorithm provide a satisfactory completed matrix. Figure-2 shows the predicted classes by the SVT algorithm at ω = 20, ω = 100 and ω = 300 respectively. We plotted the one predicted using neural network method in red color for comparison purposes. We tried different values of iterations also to get the best prediction accuracy. From the graphs one can find that at small value of ω such as ω = 20, the completed matrix is polluted with very small values, and there is not much difference in the values of three classes which is very difficult normalize. Also, for any values of ω > 100 there is not much difference in the predicted values. We found ω = 300 with 1800 iterations as most satisfactory values for parameters to achieve the best results in for this data matrix.


Actual
Predicted
Predicted class values when ω = 20
First 100 classes


Actual
Predicted
Predicted class values when ω = 100
First 100 classes

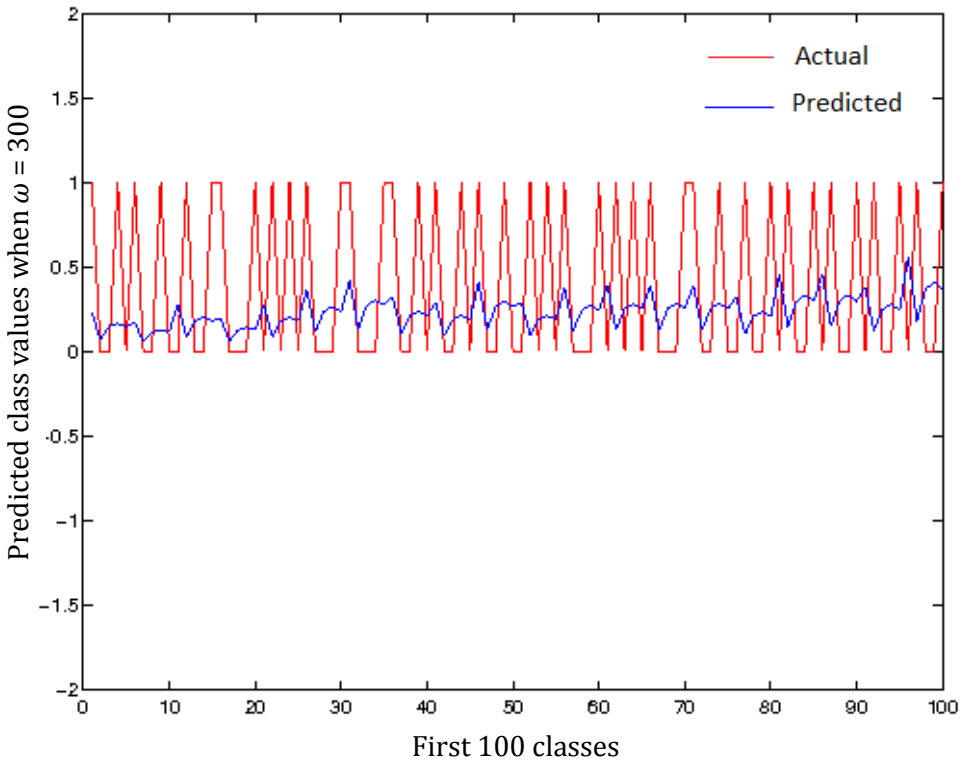


Figure-2: The predicted class values using SVT

As indicated above, to check the accuracy of the prediction by SVT we compared our results with the results generated by neural network on the same data. The class values in the results generated by neural network were either 0's or 1's. Where, 1 indicates the predicted class of the protein secondary structure. For comparison we normalized our predicted class values to 0's and 1's.

The normalized scores which is used as a probability value were calculated as,

$$p_i = \frac{score_i}{\max(all\ class) + \min(all\ class)}$$

Where class is the column congaing the predicted value of protein secondary structure and *i* indicates a value in a class column of the matrix.

We normalized these probability values to make them comparable with the neural network class values as,

$$p_i = \begin{cases} 1 & if\ class\ A_i > class\ B_i\ and\ class\ A_i > class\ C_i \\ 0 & \end{cases}$$

Where, 1 indicates the class predicted by the SVT for a particular protein.

Results and Conclusions:

Earlier our group has tested the SVT on different type of data sets like image and tournament statistics predictions, and SVT performed exceptionally well in those

cases In this project we concentrated on protein profile classification in to protein secondary structures. We tried to classify the unknown protein profiles using SVT and we compared our results with the results generated using the neural network approach. The accuracy we found for SVT to predict the protein secondary structure classes as Coils is 42.65%, as β −sheet is 76.5 % and as α −helix is 44.62% which is very low as compared to other prediction models.

This indicates that there is a lot of space to improve our predictive method. For example, conversion of available feature values in each profile to larger or smaller values can be beneficial in some cases. Also the approach for the normalization of results can be changed in many ways to get some improvements in the results. There can be many improvements in the implementation of the algorithm also.

References:

[1] Qian, N., & Sejnowski, T. J. (1988). Predicting the secondary structure of globular proteins using neural network models. Journal of Mol. Biol., 202, 865–884.

[2] Rost, B., & Sander, C. (1993). Prediction of protein secondary structure at better than 70% accuracy. Journal of Molecular Biology, 232, 584–599.

[3] Altschul, S. F., Madden, T. L., Schaeffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research, 25, 3389–3402.

[4] Jones, D. (1999). Protein secondary structure prediction based on position-specific scoring matrices. Journal of Molecular Biology, 292, 195–202.

[5] Cuff, J. A., & Barton, G. J. (2000). Application of multiple sequence alignment profiles to improve protein secondary structure prediction. Proteins: Structure, Function and Genetics, 40, 502–511.

[6] J. Abernethy, F. Bach, T. Evgeniou, and J.P. Vert, Low-rank matrix factorization with attributes, Arxiv preprint cs/0611124, (2006).

[7] Y. Amit, M. Fink, N. Srebro, and S. Ullman, Uncovering shared structures in multiclass classification, in Proceedings of the 24th international conference on Machine learning, ACM, 2007, pp. 17–24.

[8] A. Argyriou, T. Evgeniou, and M. Pontil, Multi-task feature learning, Advances in Neural Information Processing Systems, 19 (2007), pp. 41–48.

[9] M. Mesbahi and G. P. Papavassilopoulos, On the rank minimization problem over a positive semidefinitelinear matrix inequality, IEEE Transactions on Automatic Control, 42 (1997), pp. 239–243.

[10] C. Tomasi and T. Kanade, Shape and motion from image streams under orthography: a factorization method, International Journal of Computer Vision, 9 (1992), pp. 137–154.

[11] Candès, Emmanuel J., and Benjamin Recht. "Exact matrix completion via convex optimization." Foundations of Computational mathematics 9.6 (2009): 717-772.

[12] Recht, Benjamin, Maryam Fazel, and Pablo A. Parrilo. "Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization." SIAM review 52.3 (2010): 471-501.

[13] Cai, Jian-Feng, Emmanuel J. Candès, and Zuowei Shen. "A singular value thresholding algorithm for matrix completion." SIAM Journal on Optimization20.4 (2010): 1956-1982