ISSN 2722-2039

Protein Structure Prediction Using Robust Principal Component Analysis and Support Vector Machine

Nur Aini Zakaria [a,1], Zuraini Ali Shah [a,2], Shahreen Kasim [b,3,*]

[a] Department of Software Engineering, Faculty of Computing, Universiti Teknologi Malaysia, 81310 Johor Bharu, Johor, Malaysia
[b] Faculty of Computer Science and Technology, Universiti Tun Hussein Onn Malaysia, 86400 Parit Raja, Johor, Malaysia.
[1] aainizakaria@gmail.com; [2] aszuraini@utm.my; [3] shahreen@uthm.edu.my
* corresponding author

ARTICLE INFO

Article history
Received April 25, 2020
Revised April 30, 2020
Accepted May 11, 2020

Keywords:
protein structure prediction
RPCA
robust principal component analysis
support vector machine

ABSTRACT

Existence of bioinformatics is to increase the further understanding of biological process. Proteins structure is one of the major challenges in structural bioinformatics. With former knowledge of the structure, the quality of secondary structure, prediction of tertiary structure, and prediction function of amino acid from its sequence increase significantly. Recently, the gap between sequence known and structure known proteins had increase dramatically. So it is compulsory to understand on proteins structure to overcome this problem so further functional analysis could be easier. The research applying RPCA algorithm to extract the essential features from the original high-dimensional input vectors. Then the process followed by experimenting SVM with RBF kernel. The proposed method obtains accuracy by 84.41% for training dataset and 89.09% for testing dataset. The result then compared with the same method but PCA was applied as the feature extraction. The prediction assessment is conducted by analyzing the accuracy and number of principal component selected. It shows that combination of RPCA and SVM produce a high quality classification of protein structure

1. Introduction

The functional and structural annotation of protein domain is one of the important roles in bioinformatics. In this context, protein structure information plays an important information key of their structural part also the features related to the biological function (S.S. Sahu et al., 2009) such as prediction of DNA binding site, implementation of a heuristic approach to find tertiary structure, reduction of conformation search space and also characterizing the folding type of a protein or its domain. S. Zhang et al. (2012) state that the exponentially growth of newly discovered protein sequences by different scientific community caused a large gap between the number of sequence-known and the number of structure-known proteins. Hence, there exist critical challenges to develop automated method for fast and accurate determination of the structures of proteins in order to reduce gap. Therefore, there is a compulsory to implement reliable and effective computational methods for identifying the structural class of newly discovered protein based on their primary sequences.

2. Objectives

The purposes of this research are: 1. To implement Robust Principal Component Analysis (RPCA) to determine the number of principal component. 2. To implement Support Vector Machine (SVM) for protein structure classification. 3. To evaluate the performance of RPCA and SVM based on accuracy

3. Methodology

Firstly, the current issues of protein structure prediction are investigated followed by collecting research materials such as journals, articles, conference paper and others. The data preprocessing conducted to gain higher and better prediction success rate and system performance. It also help to minimizing error in preparation be validated by machine learning algorithm. Datasets by Ding and Dubchak (2012) filtered to remove unnecessary values and information. Research continues by applying Principal component analysis (PCA) and RPCA (Croux and Ruiz-Gazen, 2005)) algorithm to extract the essential features from the original high-dimensional input vectors. The process continued by experimenting SVM with RBF kernel using the reduced and normalized features by PCA and RPCA. The final phase is the prediction assessment of the application of RPCA and SVM by the comparison of recognition ratio compared between different methods and methods used by previous researcher. Performance testing of this research by comparing classification result of protein by overall accuracy that expressed in equation 1.

$$correctly\ recognize\ protein = \frac{correctly\ recognize\ number\ of\ query\ protein}{total\ number\ of\ protein} \quad (1)$$

4. Result and Discussion

The experiment was conducted by using three approaches in order to analyse the performance of RPCA and SVM. In order to gives a clear view on performance of RPCA, the method was compared with the PCA (the basic of RPCA) and SVM. In order to select the components that contain >60% of variance, the number of PC selected are different accordingly. Table 1 shows that number of selected PC in training dataset is lower compared to testing dataset. Table 2 shows the accuracy percentage of tested approach divided by training and testing datasets.

Table 1. Number of PC selected for classification

Feature extraction	Number of PC selected for classification	
	Training	**Testing**
PCA	2	3
RPCA	2	4

Table 2. Comparison of SVM, PCA + SVM and RPCA + SVM

Technique	Training Dataset Accuracy (%)	Testing Dataset Accuracy (%)
SVM	84.25	84.16
PCA+SVM	74.79	84.68
RPCA+SVM	84.41	89.09

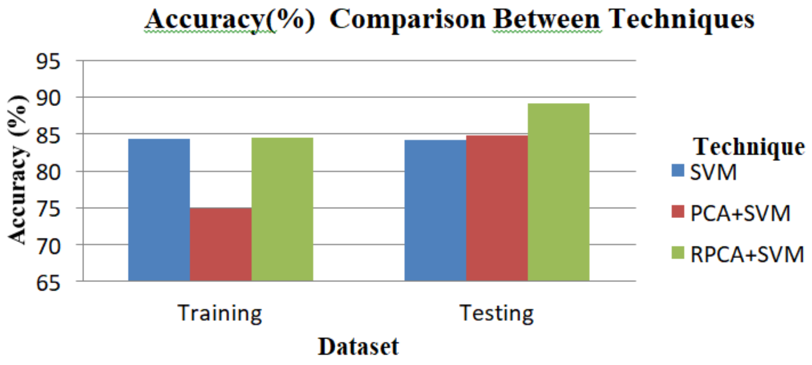

Fig 4. Accuracy comparison between techniques.

Based on this analysis, it can be assume that difference between data characteristic will influence the number of sufficient PCs required in both PCA and RPCA approach. Number of PC requires for training dataset is less then testing dataset since the size of training dataset is larger so it may contain higher information and better interpretation on features compared to testing dataset

From the results in Figure 1 it can be seen that non-extracted features technique (only SVM) gain a high percentage of accuracy (84.25% and 84.16%). However, the result can be doubt since models built on extracted features may be of higher quality, because the data is described by fewer, with more meaningful attributes. Results obtain by combination of PCA and SVM is 74.79% on training dataset and 84.68% on testing dataset. The accuracy on both datasets is quite high but still lower than combination of RPCA and SVM technique (84.41% on training dataset and 89.09% on testing dataset). The gap seems to be higher in training dataset may because of the larger number of outliers. RPCA seems to perform the best since this method do not influenced much by outliers and its ability to detect exact fit situation.

Table 3 shows the comparison of accuracy percentage of PCA and RPCA combination with SVM. Even according by number of component, the RPCA method always seems to lead in terms of accuracy. This proves the effectiveness of RPCA approach. Table 3 also shows the increasing pattern of the accuracy for both datasets. It can be assume that higher number of PC contain much more data information lead to higher accuracy.

Table 3. Comparison of PCA+SVM and RPCA+SVM based on number of component

Number of Principal Component (PC)	Accuracy (%) for Training Dataset		Accuracy (%) for Testing Dataset	
	PCA +SVM	RPCA+SVM	PCA+SVM	RPCA+SVM
1	51.91	80.60	55.84	77.92
2	74.79	84.41	84.68	84.94
3	82.75	86.90	87.53	88.05

L. Singh, G.Chetty and D.Sharma (2012) apply the same dataset (feature vector described by Ding and Dubchak, 2001) to predict protein structure using PCA and LDA based in Extreme Learning Machine (ELM). According to the Table 4, it can be seen that proposed method used in this research shows promising results in term of accuracy obtained compare to the proposed method proposed by L. Singh, G.Chetty and D.Sharma (2012). This shows that feature extraction using

RPCA and classification using SVM is an efficient method for protein structure prediction. It also shows that method proposed by L. Singh, G.Chetty and D.Sharma (2012) has drawbacks in due to the outliers and low ability in detection of exact fit situation.

Table 4. Accuracy comparison between method

Method	Accuracy (%)
LDA-ELM	77.67
PCA-ELM	82.45
RPCA-SVM	89.09

5. Conclusion

This research focus is on protein structural classification. Protein Structure classification is important for identification of protein function. As the protein structure classification n is a first and key step in protein structure prediction, it becomes an increasingly challenging task. Recently, the exponentially increase of sequence data protein cause the increasing of the requirements for reliable and effective computational method for protein structure classification. Protein structure classification is very important in bioinformatics field. Proposed feature extraction method, Robust Principal Component Analysis (RPCA) combines with Support Vector Machine (SVM) shows that data with extracted features can obtain higher accuracy (84.41% for training dataset and 89.09% for testing dataset). It also shows that RPCA works well with highly corrupted data especially dataset with outliers.

References

[1] Croux, C. and Ruiz-Gazen, A. (2005), "High breakdown estimators for principal components: the Projection-pursuit approach revisited", Journal of Multivariate Analysis, 95, 206-226

[2] Ding, Chris HQ, and Inna Dubchak. (2001), "Multi-class protein fold recognition using support vector machines and neural networks." Bioinformatics 17.4: 349-358.

[3] Singh, Lavneet, Girija Chetty, and Dharmendra Sharma.(2012) "A novel approach to protein structure prediction using PCA or LDA based extreme learning machines." Neural Information Processing. Springer Berlin Heidelberg.

[4] Li L, Cui X, Yu S, Zhang Y, Luo Z, Yang H, et al. PSSP-RFE: Accurate Prediction of Protein structure by Recursive Feature Extraction from PSI-BLAST Profile, PhysicalChemical Property and Functional Annotations.” PLoS ONE 9(3): e92863, doi:10.1371/journal.pone.0092863, (2014) .