Int. Jour. of App. Sc. and Eng. (IJASE) **3**(1) : June 2015, 1-11

DOI NO. 10.5958/2322-0465.2015.00002.7

Protein Tertiary Structure Classification based on its Physicochemical property using Neural Network and KPCA-SVM: A Comparative Study

Ashwini M. Jani[1] and Kalpit R. Chandpa[2]

[1]Department of Computer Science Engineering and Information Technology, SVM Institute of Technology, Bharuch, India
[2]Department of Computer Science Engineering and Information Technology, SVM Institute of Technology, Bharuch, India

*Corresponding author: ashwini_jani@yahoo.com

ABSTRACT

Proteins are one of the most important molecules in living organisms so they play a vital structural role in the cells of living organism. They are constructed of several polypeptide chains of amino acids, which fold into complex tertiary Structure. The knowledge of the protein function is directly dependent on its three dimensional (tertiary) structure. The Physicochemical properties of proteins always guide to determine the quality of the protein tertiary structure. Therefore it has been rigorously used to distinguish native or native like structure from other predicted structure. The experiments were conducted on the CASP dataset to classify RMSD (target class) near to native protein tertiary structure or not. Kernel principal component analysis (KPCA) is used for feature extraction since it performs better than PCA on protein tertiary structure dataset due to their nonlinear structures. The proposed model compare with neural network classification method. The experiments conducted shows that support vector machine combined with KPCA feature extraction performs better than neural network classifier. More than our results show better performance in Gaussian KPCA feature extraction with respect to other kernels.

Keywords: Protein Tertiary Structure, Physicochemical Property, Data Mining, Classification, Kernel-PCA, CASP dataset, Support Vector Machine, Neural network.

Protein plays an important role in the life support. The study of protein tertiary structure contributes to protein function and also used for medicine design and drug discovery. The Physicochemical properties of amino acids and their solvent environment are the key determinants in folding a protein sequence into its unique tertiary structure [1]. Data mining is the process of discovering interesting knowledge, such as patterns, associations, changes,

anomalies and significant structures, from large amount of data stored in databases [2]. In data mining classification task is used for identifying which of a set of categories a new observation belongs, on the basis of training set of data. Classification is considered an instance of supervised learning where a training set of correctly identified observations is available [3]. Data mining classification techniques are more effective and appropriate for the classification of protein data so, there is need a mechanism which improves the problems which are faces by traditional methods. The SVM [4] is new and promising technique for Data Classification. The SVM classifier is widely used in bioinformatics due to its high accuracy, and ability to deal with high- dimensional data. Support Vector Machine (SVM) is a robust classification and regression technique that maximizes the classification accuracy of a model without over fitting the training data. The protein structural data has so many dimensions and difficult to classify based on known class labels. The support vector machine is a new classification algorithm in data mining that deal with high dimensional database and gives better classification result with high accuracy [5]. Here we used six physicochemical properties of protein tertiary structure namely total surface area, Euclidian distance, total empirical energy, secondary structure penalty, sequence length and pair number.

A technique used to extract feature is the principal component analysis (PCA) [6]. It is a way of identifying patterns in data, and expressing the data in such a way as to highlight the similarities and differences. The main advantage of PCA is that once you have found these patterns in the data and you compress the data by reducing no. of dimensions without much loss of information. However this technique reveals only linear structures in a given dataset. So the extension of this method to non linear data is Kernel PCA. The concept of this technique is to map the data into a high dimensional feature space by using kernel functions and PCA is applied on mapped data.

In this paper, we propose to classify protein tertiary structure based on its physicochemical property and Root Mean Square Deviation (RMSD) of Features. The experiments were conducted on the CASP dataset to classify RMSD (target class) near to native protein tertiary structure or not. To this end, two classification techniques are used: Proposed Method KPCA-SVM and Neural Network.

The remainder of this paper is organized as follows: Section II outlines the basic concepts involving KPCA. Section III describes the classification techniques Support vector machine and Neural Network. Section IV presents the overview of features and dataset descriptions with proposed methodology. Section V presents the experimental result that evaluates the adopted techniques performance, by the end, a final conclusion and recommendation for future work are presented.

KERNEL PCA

Principal component analysis is a mathematical technique whose purpose is to transform

a number of correlated variables into a number of uncorrelated variables called "Principal Components" [7]. The conventional PCA detects only linear structures in a given dataset. A more generalized technique has been introduced to learn the nonlinearities using kernels, so called kernel PCA (KPCA). The basic idea of KPCA is to map original dataset into high-dimensional feature space via a specific kernel functions and then apply standard PCA algorithm on it. One of the main advantages of KPCA is that by choosing a specific kernel function we can have a previous idea about the type of non linear components before extracting them [8]. As said before one of the main advantages of standard PCA towards KPCA is that we can reconstruct easily the original data by simply using the principal components.

CLASSIFICATION TECHNIQUES

In order to classification of protein tertiary structure in literature various well known classification methods are available. We have used support vector machine and neural network classification techniques. In this paper our aim is to discuss and compare classification results of both techniques.

(A) Support Vector Machines

The support vector machine (SVM) is a new and promising technique for data classification. After the development in the past five year, it has become an important topic in Data mining and machine learning. The Basic Steps of applying SVM for Classification Can be stated briefly as follows [9]:

First, map the input vectors into one feature space (possible with a higher dimension), either linearly or non-linearly, which is relevant with the selection of the kernel function. Then, within the feature space from the first step, seek an optimized linear division, i.e. construct a hyper plane which separates two classes (this can be extended to multi-class).SVM training always seeks a global optimized solution and avoids over fitting , so it has the ability to deal with a large number of features.

Data: $<\mathbf{x}_i, y_i>$, i=1,..,l

$\mathbf{x}_i \varepsilon R^d$

$y_i \in \{-1,+1\}$

Hyperplane: $\mathbf{x}_i \bullet \mathbf{w}+b \geq +1$ when $y_i = +1$

$\mathbf{x}_i \bullet \mathbf{w}+b \leq -1$ when $y_i = -1$

Now the classifying function will have the following form:

$$f(x) = (\Sigma \alpha_i x_i)\, x + b \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (1)$$

It is called linear discriminate function. SVM finds a linear separating hyperplane with maximal margin in the higher dimensional space.

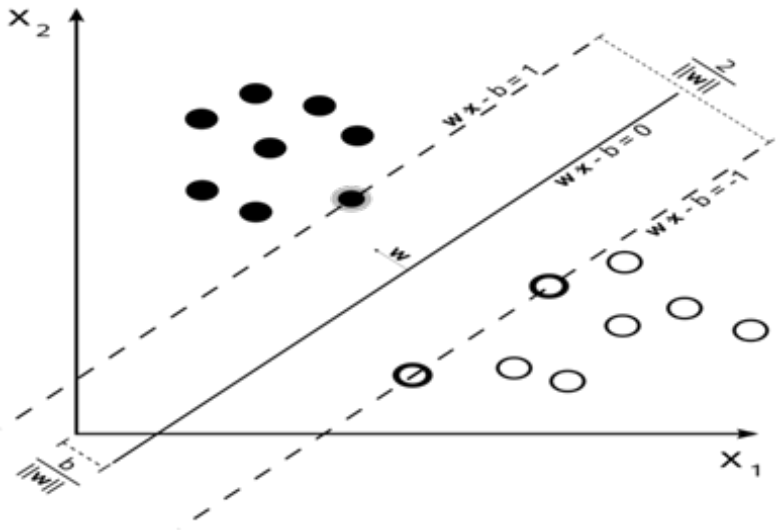

Figure 1: Basic Concept of SVM

In real world it is not likely to an exactly separate line dividing the data within the space, and we might have a curved decision boundary. It is better for the smooth boundary to ignore few data points than be curved or go in loops, around the outliers. This can be handling by slack variables. In this case w and b can be finding by following equations:

$$\phi(\omega) = \text{minimize } \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}\xi_i$$

The dual problem for soft margin classification:

Find $\alpha_1 \ldots \alpha_N$ such that

$Q(\alpha) = \Sigma\alpha_i - \frac{1}{2}\Sigma\Sigma\alpha_i\alpha_j y_i y_j x_i^T x_j$ is maximized and

(1) $\Sigma\alpha_i y_i = 0$

(2) $0 \le \alpha_i \le C$ for all α_i

The discriminate function becomes:

$$f(x) = \Sigma\alpha_i y_i x_i^T x + b \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (2)$$

The mathematical function used for the best margin is known as kernel function. SVM supports the linear, polynomial, radial basis function (RBF) and sigmoid kernel function. The data points that lie on the margins are known as support vectors. The margin will be wider between the two categories for better model and classification of new records. The parameter C is used as highly sensitive parameter which determines the flexibility of the margin of hyperplane. The most important thing of employing SVM is to select a suitable kernel function. The most basic kernel functions are categorised into four categories:

- Linear: $K(x_i,x_j)=x_i^T x_j$
- Polynomial: $K(x_i,x_j)=(\gamma x_i^T x+r)^d$, $\gamma>0$
- Radial Basis Function : $K(x_i,x_j)=\exp(-\gamma||x_i^T x_j||^2,\gamma>0$
- Sigmoid: : $K(x_i,x_j)= \tanh(\gamma x_i^T x_j+r)$

Where γ,r and d are kernel parameters.

Now for soft margin the discriminate function becomes:

$f(x) = \Sigma \alpha_i y_i K(xi,xj) + b$ ..... (3)

Neural Network

In this research we use multilayer perceptron as a three layer feed forward network with weight adjusted by conjugate gradient minimization factor. We used Back propagation learning algorithm for performing supervised learning and training of MLP. The BPN Algorithm uses a gradient search technique to minimize a cost function equivalent to the MSE between actual and desired network outputs. The figure shows back propagation network architecture. The BPN Algorithm propagates back the error between the desired and network output through network. After providing an input pattern the network output is compared with target pattern and error of each output unit is calculated. The performance can be improved and the occurrence of local minima reduced by allowing extra hidden layers, lowering the gain term, and by training with different initial random weights [10].

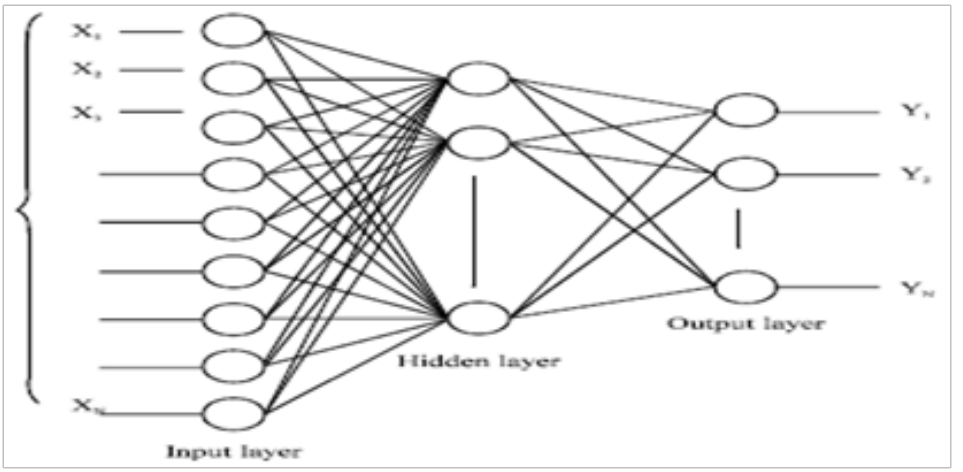

Figure 2: Back propagation Network Architecture

MATERIALS AND METHODS

In order to implement task of classifying tertiary structure data, there is requirement of some benchmark datasets including relevant features and target class. This section introduces those datasets description and proposed methodology.

(A) Dataset Description

The dataset that we have chosen for classifying protein tertiary structure is Physicochemical Properties of Protein Tertiary Structure. There are a total of 95091 modeled structures of 4896 native targets. The modeled structures are taken from protein structure prediction center (CASP-5 to CASP-9 experiments), public decoys database and native structure from protein data bank (RCSB). The dataset used in the study is available as supplement at *http://bit.ly/RF-PCP-DataSets*. The Table 1 shows dataset description.

Table 1. Dataset Description

Data Set Characteristics:	Multivariate	No. of instances:	95091
Attribute Characteristics:	Real	No. of Attributes:	9
Associated tasks:	Classification/ Regression	Data Denoted:	2013
Missing Values:	N/A	No. of Classes	3

(B) Proposed Methodology

The main objective of proposed system is to classify protein's tertiary structure near to its native or predictive structure with RMSD target class using KPCA-SVM classification method. For data preprocessing we will apply feature extraction method KPCA for optimization of relevant features from dataset. Since KPCA is extension of principal component analysis and can reveal nonlinear kernel principal components that are more appropriate to complex and nonlinear data such as protein structures. Figure 3 shows proposed system model.

Sample Dataset: The dataset that we have chosen for classifying protein tertiary structure is Physicochemical Properties of Protein Tertiary Structure. Physicochemical properties of proteins always guide to determine the quality of the protein structure; therefore it has been used to distinguish native or native like structure from other predicted structures.

Feature Selection: Here, six physicochemical properties namely total surface area (Area), Euclidean distance (ED), total empirical energy (Energy), secondary structure penalty (SS), sequence length (SL) and pair number (PN) are selected. These features are selected based on expert advice and their importance in protein tertiary structure. The feature selection makes the classification model efficient and accurate.

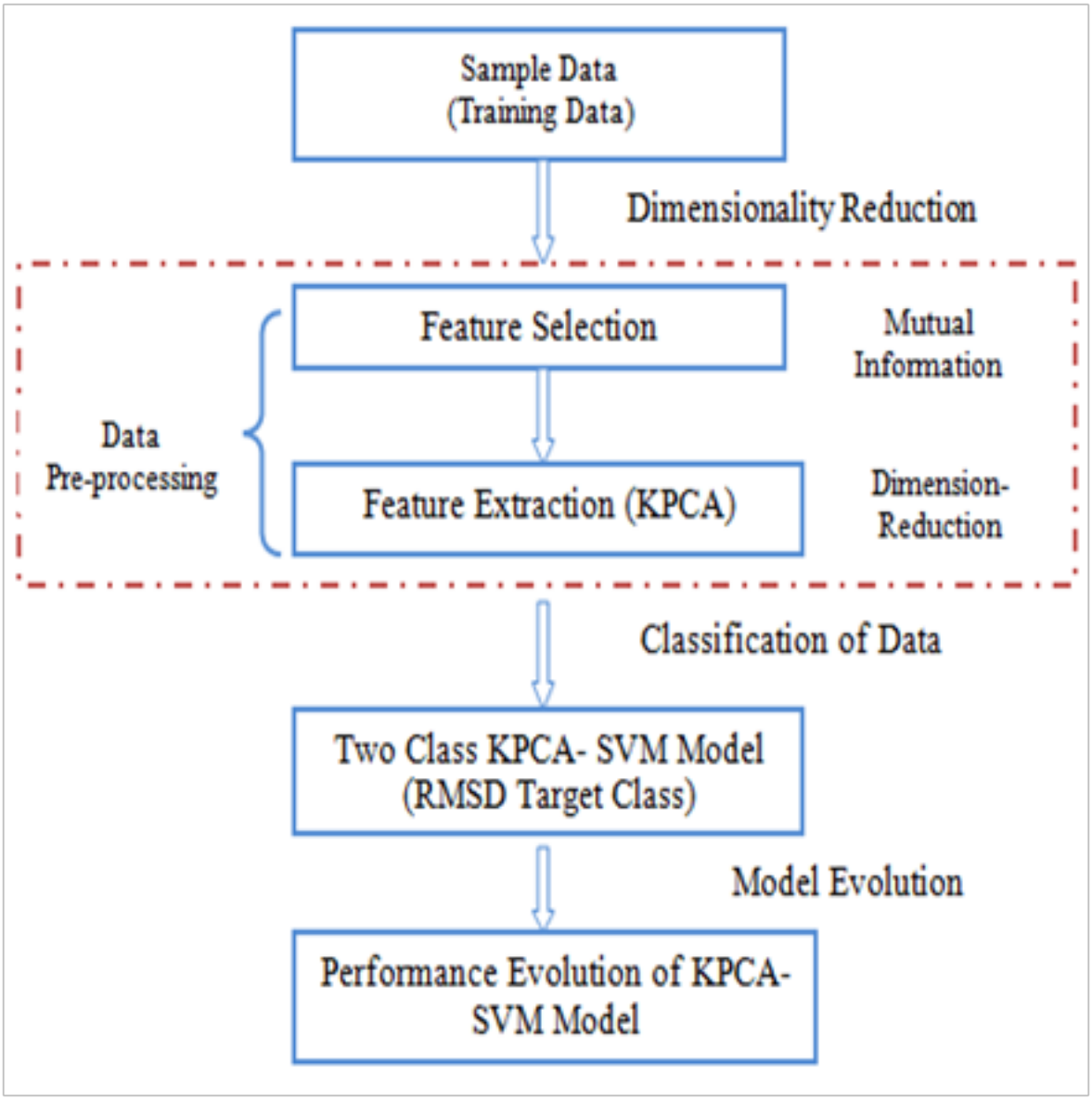

Figure 3: Proposed System Model

Feature Extraction: It is the special form of dimensionality reduction. It combines the attributes into a new reduced set of features. When the input data to the algorithm is too large the data will be transformed into reduced representation set of features. The main advantage of feature extraction is it reduces the measurement, storage requirements and also reduces the training

and utilization times of the final model. Here we have used kernel principal component analysis method for feature extraction. It is extension of principal component analysis with various kernels for non-linear data mapping into hyper plane.

Model Evaluation: The Proposed System consists of a unified solution to classify protein tertiary structure. we have implemented support vector machine with various kernel functions with modified parameter values as per data and classify protein tertiary structure based on their physicochemical property. With selected features we have classified RMSD (Root Mean Square Deviation) target class for native tertiary structure and predictive tertiary structure. For improving Classification performance we have applied Kernel principal component analysis (KPCA) feature extraction method. At the last phase of result analysis we will compare our proposed approach with neural network classifier based on their classification accuracy/ RMSE and training time.

RESULTS

The experiment has been carried out on above mentioned dataset in MATLAB R2014a Software tool. This section shows implementation results for taken datasets with respect to parameters such as Classification performance and Training time.

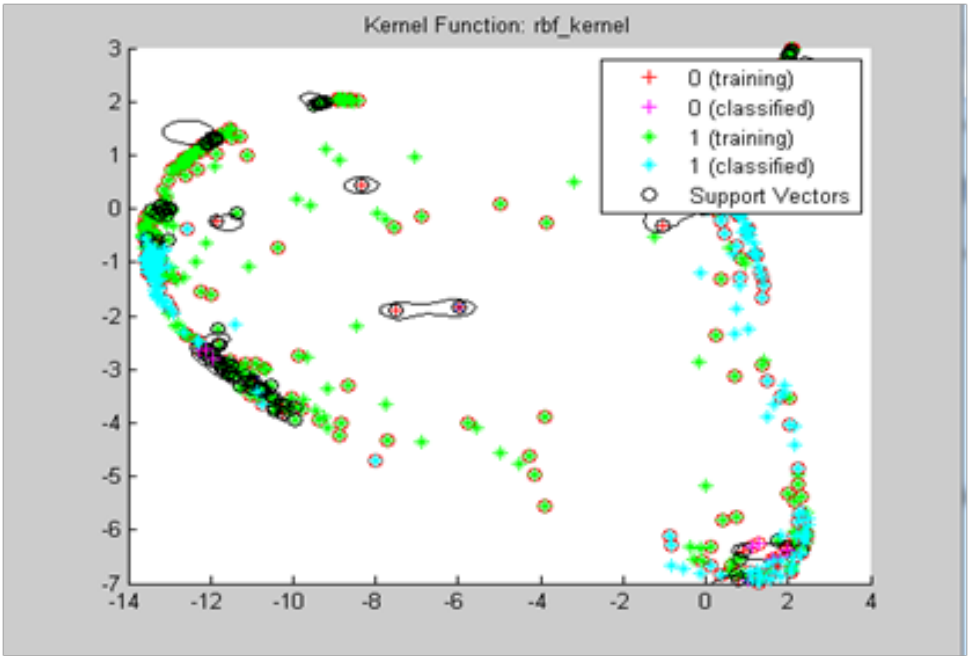

Figure 4: KPCA- SVM Classifier with RBF kernel Function

Our target is to optimize the classifier performance and to reach a higher percentage in the accuracy. The inputs are training data, the kernel parameter sigma and the regularization constant C. Each pair of the kernel argument and its relevant parameters called Model. Here we have used all the four kernel parameters namely linear, sigmoid, polynomial and radial basis function with their parameter values. When using two class SVM we had found that the SVM classifier with RBF kernel function presents better results than other kernel functions. Figure 5 shows classification performance of KPCA-SVM with various kernel functions.

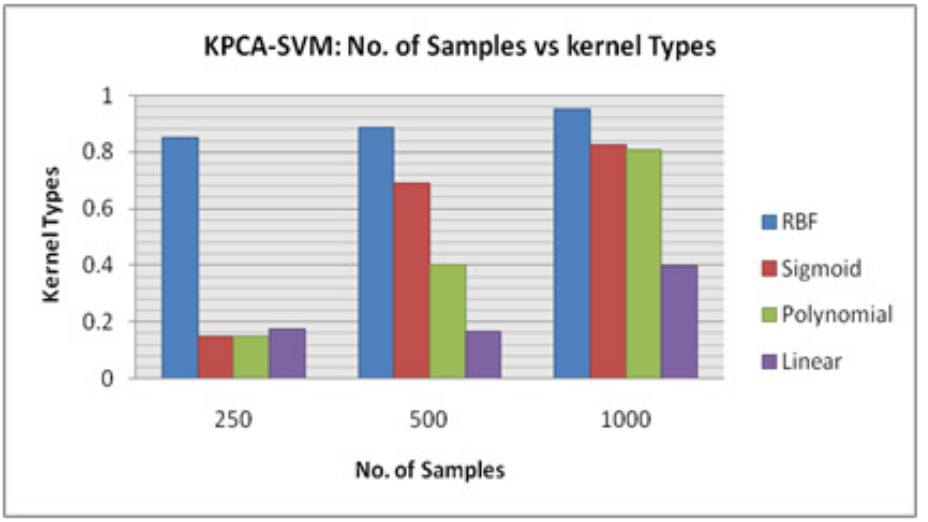

Figure 5: KPCA-SVM: No. of Samples vs. Kernel Types.

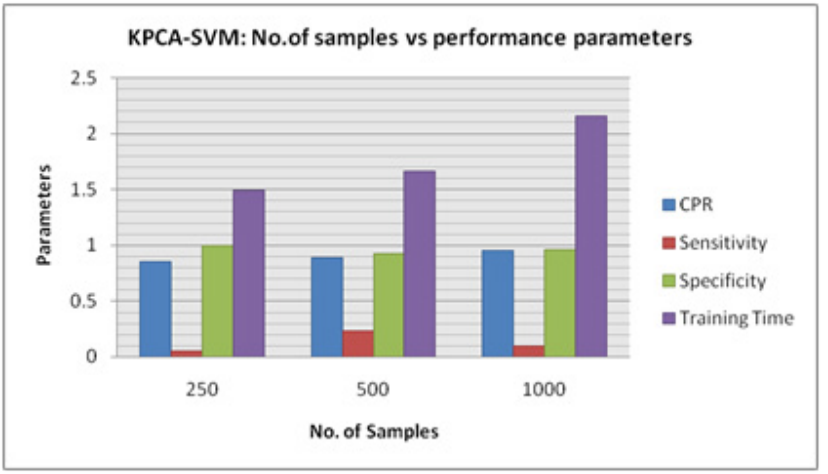

Figure 6. KPCA-SVM: No. of Samples vs. Performance parameters

Figure 6 shows performance of two-class KPCA-SVM with various Performance parameters. The graph shows that two-class KPCA-SVM classifier classifies the data with Class performance rate of 95 % and with small training time of approx 2 seconds. The performance of the classifier is increase with the number of samples increased.

Model Validation with Artificial Neural Network

We used Multilayer Perceptrons (MLP) as a Network model for SVM Model validation. The BPN Algorithm uses a gradient search technique to minimize a cost function equivalent to the MSE between actual and desired network outputs. In this section we have implemented ANN for classification of taken dataset. For training performance we have used a Root Mean Square error which is used for estimate the errors. If the RMSE is less, than it shows that performance is better. So, we have taken the Average training performance of all three data samples. The graph shows comparison result analysis of artificial neural network with KPCA-support vector machine. As shown in graph as no. of samples increase RMSE of ANN also increased and classification performance decrease. But in case of KPCA support vector machine as no. of samples increase the classification performance is also increased.

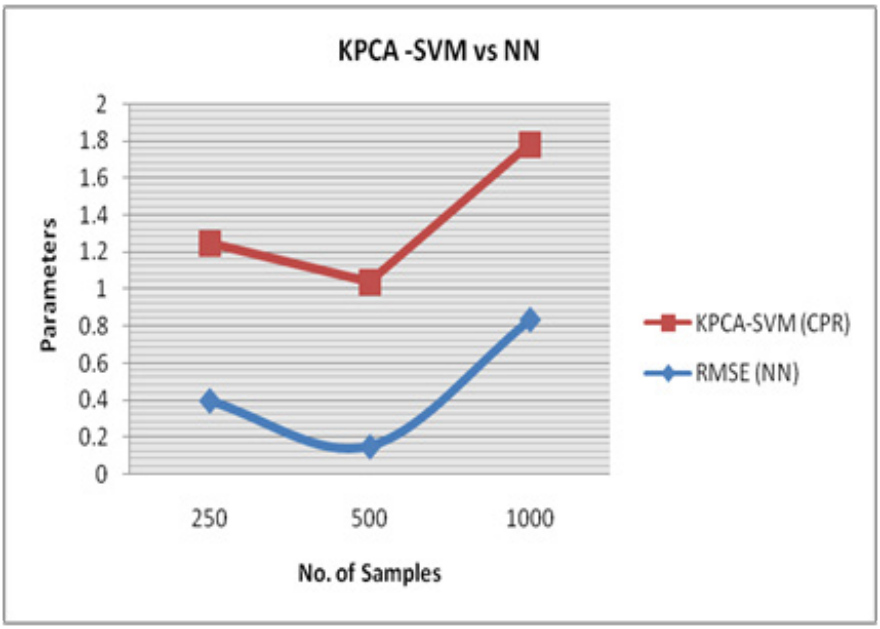

Figure 7. KPCA-SVM vs. NN

CONCLUSION

The protein tertiary structure classification using recent technologies is complicated and time consuming process and these techniques are also expensive. The classification of protein data

according to their structure and function is recent research topic and Support Vector Machine algorithm gives high accuracy of classification for biological database. We also concluded that Relevant features extracted from the relevant feature extraction method gives to the classifier for improving classification accuracy and reducing training time.

The experiment has been conducted on the CASP dataset which includes physicochemical property of protein tertiary structure. Experimental result shows that Two-class KPCA-SVM classifies data with class performance rate of 95 % and with minimum training time approx 2 seconds for binary classification of RMSD target class. The result analysis shows that KPCA-SVM with radial basis kernel function for parameters C=100 and rbf_sigma= 0.1 gives best classification accuracy as compare to another kernel functions.

In our future work, we intend to enlarge more than one no. of classes for multiclass classification purpose. Multiclass is the problem of classifying instances into more than two classes. SVM classifier permits use of more than two classes in non-linear separable case that is extension of binary classification. For non linear case we are going to use RBF kernel function for best data transformation of learning purpose. We have another two classes named TM- score (Target Modeling) and GDT-score for the same features of CASP dataset. For multiclass purpose we can use two methods one-against-one and all-against-all. For improving classification performance we can apply kernel principle component analysis feature extraction method because it performs better than principle component analysis and gives better classification performance with less training time.

REFERENCES

Prashant Singh Rana, Harish Sharma, Mahua Bahattacharya and Anupam Shukla, "Quality assesment of modelled protein stucture using physicochemical properties" *Journal of Bioinformatics and Computational Biology* © Imperical College Press.

Li,J,;Wong, L. and Yang, Q. 2005. Data Mining in bioinformatics, IEEE Intelligent System, IEEE Computer Society.

Kalid Raza "Application of Data Mining in Bioinformatics" *Indian Journal of Computer Science and Engineering* **1**(2): 114-118.

"Application of Support Vector Machines in Bioinformatics" .pdf by jung-Ying Wang.

C.J.C. Burges. 1998. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, (2): 121–167.

I.T. Jolliffie, Principal component analysis, 2nd ed. Spinger, Oct.2002.

L.I.Smith, 2007. "A Tutorial on Principal Component Analysis Introduction", Statistics, 98-98.

D.Patra, M. K. Das, and S. Pradhan 2010. "Integration of FCM, PCA and neural networks for classification of ECG arrhythmias," *IAENG International Journal of Computer Science*.

A user's guide to support vector machine.pdf Asa Ben-Hur,Jason Weston.

Mehdi Khashei, Ali Zeinal Hamadani and Mehdi Bijari 2014. "Neural network and SVM Classifiers accurately predict lipid binding proteins, irrespective of sequence homology", Elsevier, Science Direct, *Journal of Theoretical Biology*.