ISSN:0975-9646

Improved Protein Function Classification Using Support Vector Machine

Ravdeep Singh
Student of M.Tech
Department of CSE
LLRIET, Moga

Prof. Rajbir Singh
Associate Prof. & Head
Department of IT
LLRIET, Moga

Dheeraj Pal Kaur
Assistant Prof. (ECE)
Department of ECE
LLRIET, Moga

Abstract-Machine learning technique is introduced as a method for the classification of proteins into functionally distinguished classes. Protein function classification is one of the most important problems in modern computational biology. Studies are conducted on a number of protein classes including RNA-binding proteins; protein homodimers, proteins responsible for drug absorption, proteins involved in drug distribution and excretion, and drug metabolizing enzymes. Support vector machine in drug discoveries and designing based on two data sets of protein with their various physic-chemical properties. Some binary classification problems do not have a simple hyper plane as a useful separating criterion. For those problems, there is a variant of the mathematical approach that retains nearly all the simplicity of an SVM separating hyper plane. The mathematical approach using kernels relies on the computational method of hyper planes. All the calculations for hyper plane classification use nothing more than dot products. Therefore, nonlinear kernels can use identical calculations and solution algorithms, and obtain classifiers that are nonlinear. The resulting classifiers are hyper surfaces in some space S. In our research, 100 different sequences of proteins are generated with two datasets, which are then classified with their physico-chemical properties independently. In addition, one two dimensional class points inside the unit disk is generated and another class of points in the annulus from radius 1 to radius 2. Then a classifier is generated on the basis of the data with the Gaussian radial basis function kernel. The default linear classifier is obviously unsuitable for this problem, since the model is circularly symmetric. Our results in the form of graphs have shown the non-linear classification of two type of data sets and then with use of curve fitting tool various other parameters are shown up which distinguish the two data sets accurately.

General Terms
Protein Database, ExPASy webserver, physico-chemical properties

Keywords
Protein function Classification, Support Vector Machine

1. INTRODUCTION

Proteins are large, complex molecules that play many critical roles in the body. They do most of the work in cells and are required for the structure, function, and regulation of the body’s tissues and organs. Proteins are made up of hundreds or thousands of smaller units called amino acids, which are attached to one another in long chains. There are 20 different types of amino acids that can be combined to make a protein.

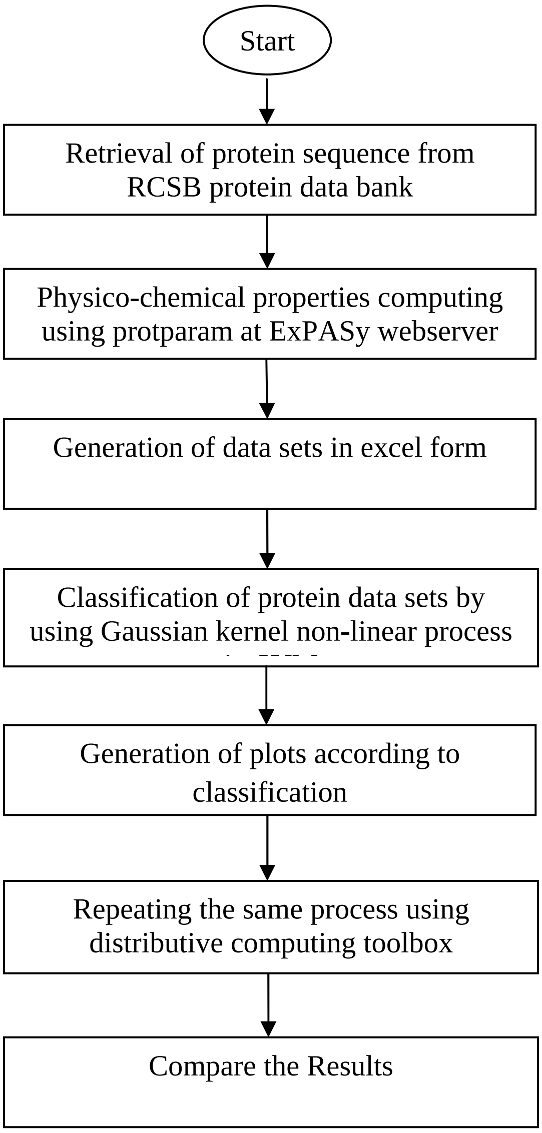
Support Vector Machines is based on statistical learning theory and is tremendously being used in data mining. The central idea is to non-linearly map the data set into a high dimensional space and use a linear discriminator to classify the data into two subsets. Its success has been demonstrated in the areas of regression, classification, and decision tree construction. The SVM classifier is widely used in bioinformatics (and other disciplines) due to its highly accurate, able to calculate and process the high-dimensional data such as gene expression and flexibility in modeling diverse sources of data .SVMs belong to the general category of kernel methods. A kernel method is an algorithm that depends on the data only through dot-products. When this is the case, the dot product can be replaced by a kernel function which computes a dot product in some possibly high dimensional feature space. This has two advantages: First, the ability to generate non-linear decision boundaries using methods designed for linear classifiers. Second, the use of kernel functions allows the user to apply a classifier to data that have no obvious fixed-dimensional vector space representation. The prime example of such data in bioinformatics are sequence, either DNA or protein, and protein structure.

For improved protein function classification we have employed SVM toolbox and distributive computing toolbox for our process. SVM has been used for the classification purpose whereas distributive computing toolbox used to do same process on two divided cores of processor of our machine.

2. METHODOLOGY

In the present methodology the major goal is protein function classification for that we had to start from analysis of things. We started by studying and analyzing the public protein database to find good achieved data bank. We concluded our analysis and search with RCSB (Research Collaboratory for Structural Bioinformatics) protein data bank, which is a worldwide repository of information about molecular structures, sequences, properties, 3D structures of proteins of various origins. We retrieved 100 random human proteins from the human protein database. Then we divided these sequences into 2 parts which gave rise into our 2 data sets.After Study of various web servers for computing the phsico-chemical properties of the proteins we came to know about ExPASy webserver. ExPASy is the SIB Bioinformatics Resource Portal which provides access to scientific databases and software tools in different areas of life sciences including proteomics, genomics, phylogeny, systems biology, population genetics, transcriptomics. We explored tools for proteomics and found our tool of interest as Protparam. Then we gave protparam an input in the form of protein sequences one by one and it gave us output in the form of physic-chemical properties of that proteins. We repeated the process until calculation of physic-chemical properties of all the protein sequences is achieved. Flow chart of process is show below.

2.1 Flow Chart of present work


Start
Retrieval of protein sequence from RCSB protein data bank
Physico-chemical properties computing using protparam at ExPASy webserver
Generation of data sets in excel form
Classification of protein data sets by using Gaussian kernel non-linear process
Generation of plots according to classification
Repeating the same process using distributive computing toolbox
Compare the Results

2.2 Physico-Chemical Properties

The physico-chemical properties we have calculated are:-

1. Extinction coefficients - Extinction Coefficient is a protein parameter that is commonly used in the laboratory for determining the protein concentration in a solution by spectrophotometry. It describes to what extent light is absorbed by the protein and depends upon the protein size and composition as well as the wavelength of the light.
2. Estimated half-life - The half-life is a prediction of the time it takes for half of the amount of protein in a cell to disappear after its synthesis in the cell
3. Instability index - The instability index provides an estimate of the stability of your protein in a test tube. Statistical analysis of 12 unstable and 32 stable proteins has revealed that there are certain dipeptides, the occurence of which is significantly different in the unstable proteins compared with those in the stable ones.
4. Aliphatic index - The aliphatic index of a protein is defined as the relative volume occupied by aliphatic side chains (alanine, valine, isoleucine, and leucine). It may be regarded as a positive factor for the increase of thermostability of globular.
5. Grand average of hydropathicity - The GRAVY value for a peptide or protein is calculated as the sum of hydropathy values of all the amino acids, divided by the number of residues in the sequence
6. Theoretical pI - Protein pI is calculated using pKa values of amino acids. The pKa value of Amino acids depends on its side chain. It has an important role in defining the pH dependent characteristics of a protein

2.3 Classification of protein data sets

After generation of datasets we generated an environment in matlab to achieve classification between two data sets. This involved the coding of Gaussian kernel non-linear process. After this process we imported our data sets into the matlab and then classified them using the Gaussian kernel non-linear process. Here it is important to mention that all the proteins will be classified according to individual physic-chemical property separately which means we will do classification six times according to each physico-chemical property. Our process will classify the proteins and then create plots which will define the classification on the visual basis. Then according to our objectives we used distributive computing toolbox. In this process we divide our processor into 2 equal parts and then proceed with our same classification method. The results expected are same with only time difference will be there in two processes. Then we employed the curve fitting tool for viewing our results of classification in additional way. We have calculated some more parameters and generated some more plots using this tool.

3. RESULTS AND DICUSSIONS

After classification process we came across some good results defining the differentiation of two protein datasets. We have plotted the results of SVM classification as well we have also used curve fitting tool to validate our results and to calculate some more parameters. All the plots and results we have calculated are following:

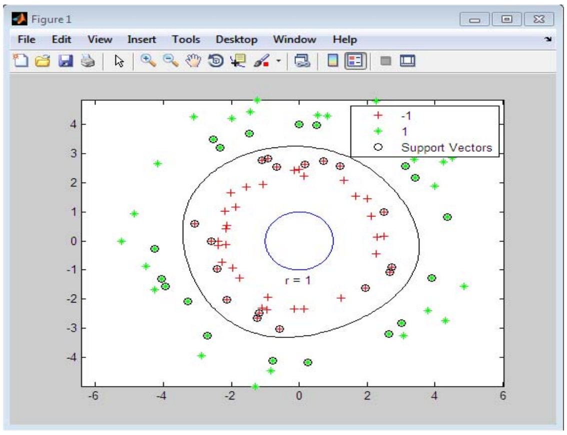


Figure 1: Plot of classification of data sets on basis of Aliphatic index.

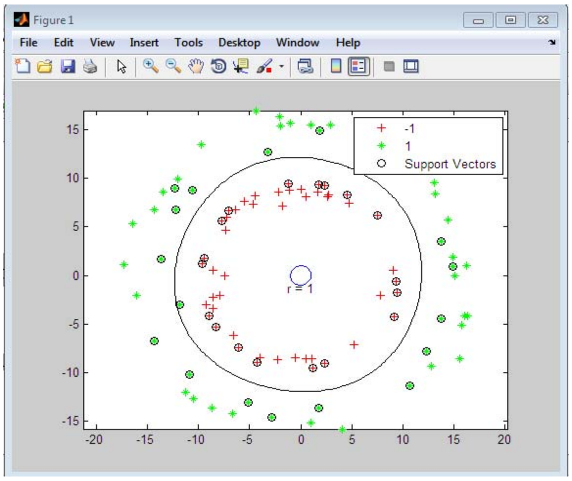


Figure 2: Plot of classification of data sets on basis of Theoretical

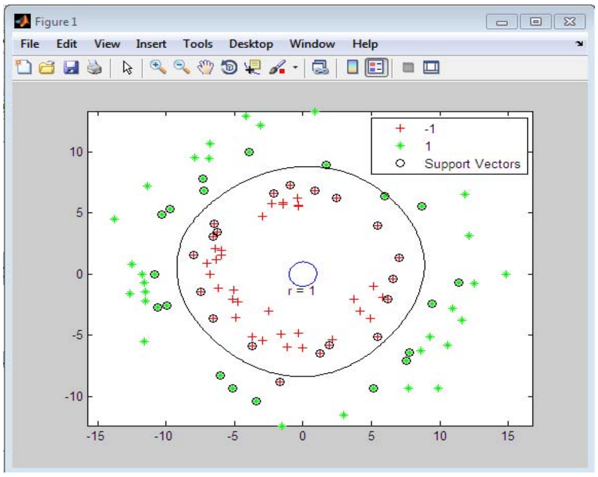


Figure 3: Plot of classification of data sets on basis of Instability index

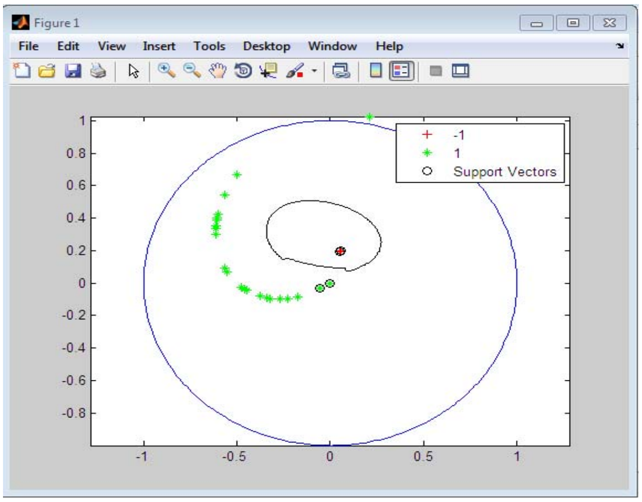


Figure 4: Plot of classification of data sets on basis of Grand average of Hydropathicity

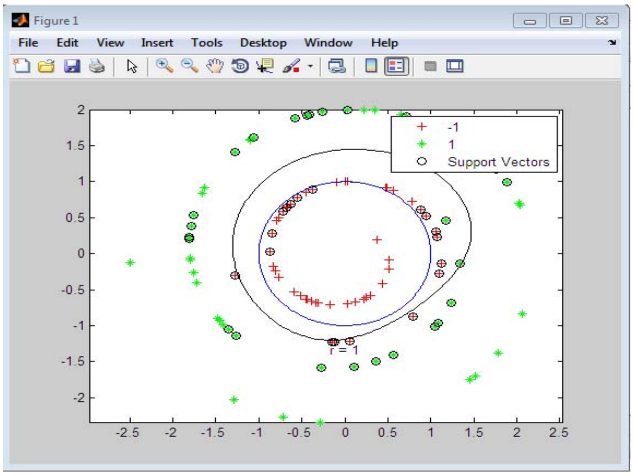
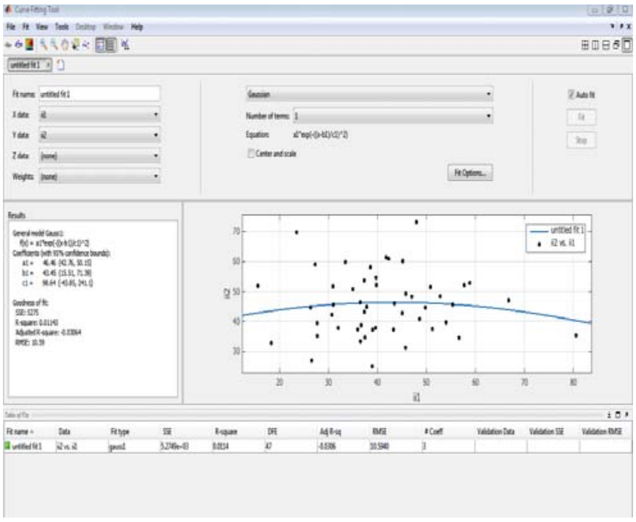


Figure 5: Plot of classification of data sets on basis of Extinction coefficients



Fit name: untitled fit 1
X data: i1
Y data: i2
Z data: (none)
Weights: (none)
Gaussian
Number of terms: 1
Equation: a1*exp(-((x-b1)/c1)^2)
☐ Center and scale
☑ Auto fit

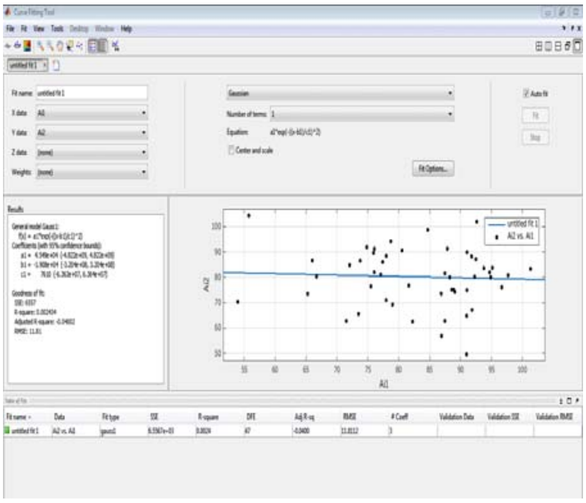
Results

```
General model Gauss1:
  f(x) = a1*exp(-((x-b1)/c1)^2)
Coefficients (with 95% confidence bounds):
    a1 =    46.46  (42.76, 50.15)
    b1 =    43.45  (15.51, 71.38)
    c1 =    98.64  (-43.85, 241.1)

Goodness of fit:
  SSE: 5275
  R-square: 0.01143
  Adjusted R-square: -0.03064
  RMSE: 10.59
```

Fit name	Data	Fit type	SSE	R-square	DFE	Adj R-sq	RMSE	# Coeff	Validation Data	Validation SSE	Validation RMSE
untitled fit 1	i2 vs. i1	gauss1	5.2749e+03	0.0114	47	-0.0306	10.5940	3			

Figure 6: Gaussian classification of Instability index



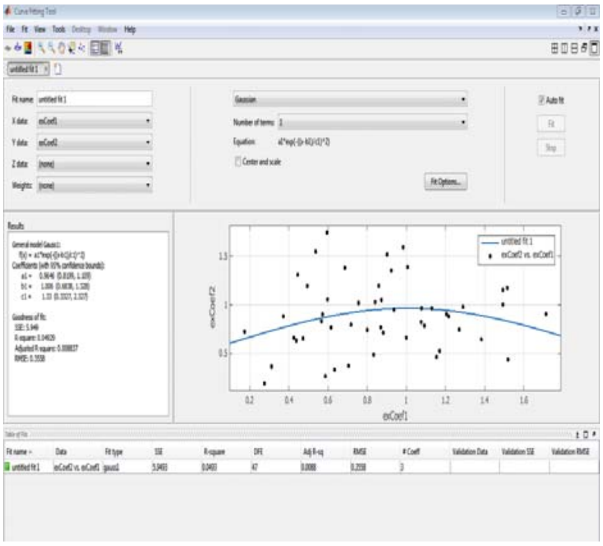
Fit name: untitled fit 1
X data: A1
Y data: A2
Z data: (none)
Weights: (none)
Gaussian
Number of terms: 1
Equation: a1*exp(-((x-b1)/c1)^2)
☐ Center and scale
☑ Auto fit

General model Gauss1:
f(x) = a1*exp(-((x-b1)/c1)^2)
Coefficients (with 95% confidence bounds):
a1 = 4.549e+04 (-4.822e+09, 4.822e+09)
b1 = -1.908e+04 (-3.204e+08, 3.204e+08)
c1 = 7610 (-6.363e+07, 6.364e+07)

Goodness of fit:
SSE: 6557
R-square: 0.002434
Adjusted R-square: -0.0402
RMSE: 11.81

Fit name	Data	Fit type	SSE	R-square	DFE	Adj R-sq	RMSE	# Coeff	Validation Data	Validation SSE	Validation RMSE
untitled fit 1	A2 vs. A1	gauss1	6.5567e+03	0.0024	47	-0.0400	11.8112	3			

Figure 7: Gaussian classification of Aliphatic index



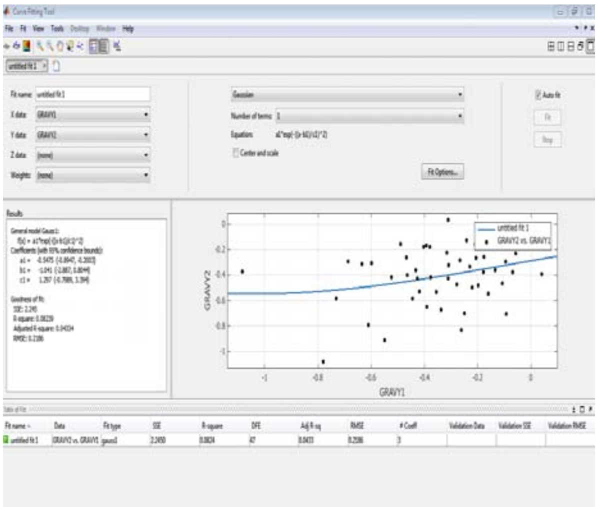
Fit name: untitled fit 1
X data: exCoef1
Y data: exCoef2
Z data: (none)
Weights: (none)
Gaussian
Number of terms: 1
Equation: a1*exp(-((x-b1)/c1)^2)
☐ Center and scale
☑ Auto fit

General model Gauss1:
f(x) = a1*exp(-((x-b1)/c1)^2)
Coefficients (with 95% confidence bounds):
a1 = 0.9646 (0.8195, 1.109)
b1 = 1.006 (0.6838, 1.328)
c1 = 1.33 (0.3327, 2.327)

Goodness of fit:
SSE: 5.949
R-square: 0.04629
Adjusted R-square: 0.008837
RMSE: 0.3558

Fit name	Data	Fit type	SSE	R-square	DFE	Adj R-sq	RMSE	# Coeff	Validation Data	Validation SSE	Validation RMSE
untitled fit 1	exCoef2 vs. exCoef1	gauss1	5.9493	0.0463	47	0.0088	0.3558	3			

Figure 8: Gaussian classification of Extinction coefficient.



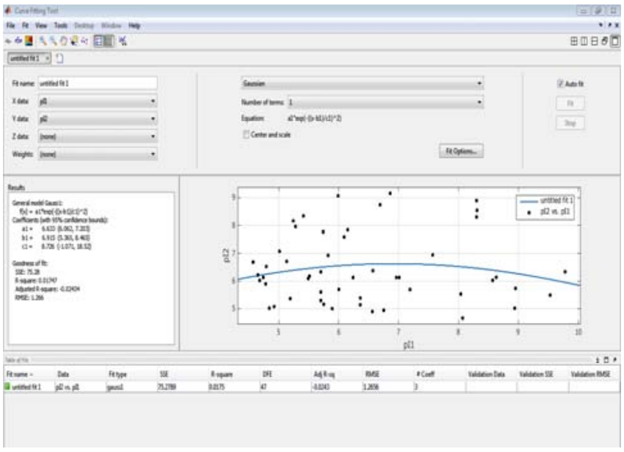
Fit name: untitled fit 1
X data: GRAVY1
Y data: GRAVY2
Z data: (none)
Weights: (none)
Gaussian
Number of terms: 1
Equation: a1*exp(-((x-b1)/c1)^2)
☐ Center and scale
☑ Auto fit

General model Gauss1:
f(x) = a1*exp(-((x-b1)/c1)^2)
Coefficients (with 95% confidence bounds):
a1 = -0.5475 (-0.8647, -0.2303)
b1 = -1.041 (-2.887, 0.8044)
c1 = 1.297 (-0.7686, 3.364)

Goodness of fit:
SSE: 2.245
R-square: 0.08239
Adjusted R-square: 0.04334
RMSE: 0.2186

Fit name	Data	Fit type	SSE	R-square	DFE	Adj R-sq	RMSE	# Coeff	Validation Data	Validation SSE	Validation RMSE
untitled fit 1	GRAVY2 vs. GRAVY1	gauss1	2.2450	0.0824	47	0.0433	0.2186	3			

Figure 9: Gaussian classification of Grand average of hydropathicity.



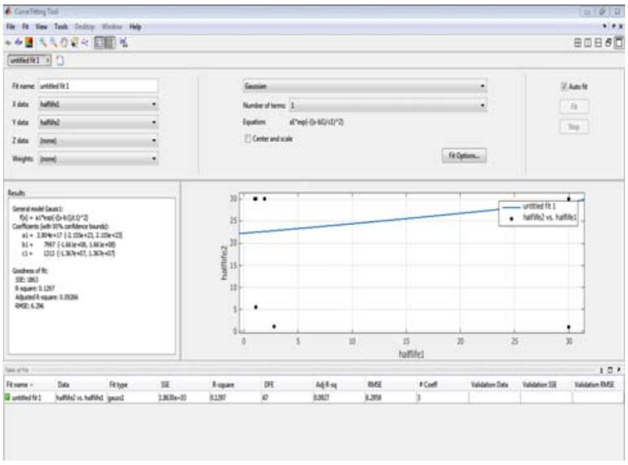
Fit name: untitled fit 1
X data: pI1
Y data: pI2
Z data: (none)
Weights: (none)
Gaussian
Number of terms: 1
Equation: a1*exp(-((x-b1)/c1)^2)
☐ Center and scale
☑ Auto fit

General model Gauss1:
f(x) = a1*exp(-((x-b1)/c1)^2)
Coefficients (with 95% confidence bounds):
a1 = 6.633 (6.062, 7.203)
b1 = 6.915 (5.365, 8.465)
c1 = 8.726 (-1.071, 18.52)

Goodness of fit:
SSE: 75.28
R-square: 0.01747
Adjusted R-square: -0.02434
RMSE: 1.266

Fit name	Data	Fit type	SSE	R-square	DFE	Adj R-sq	RMSE	# Coeff	Validation Data	Validation SSE	Validation RMSE
untitled fit 1	pI2 vs. pI1	gauss1	75.2789	0.0175	47	-0.0243	1.2656	3			

Figure 10: Gaussian classification of Theoretical



Fit name: untitled fit 1
X data: halflife1
Y data: halflife2
Z data: (none)
Weights: (none)
Gaussian
Number of terms: 1
Equation: a1*exp(-((x-b1)/c1)^2)
☐ Center and scale
☑ Auto fit

General model Gauss1:
f(x) = a1*exp(-((x-b1)/c1)^2)
Coefficients (with 95% confidence bounds):
a1 = 3.804e+17 (-2.155e+23, 2.155e+23)
b1 = -7997 (-1.661e+08, 1.661e+08)
c1 = 1313 (-1.367e+07, 1.367e+07)

Goodness of fit:
SSE: 1863
R-square: 0.1297
Adjusted R-square: 0.09266
RMSE: 6.296

Fit name	Data	Fit type	SSE	R-square	DFE	Adj R-sq	RMSE	# Coeff	Validation Data	Validation SSE	Validation RMSE
untitled fit 1	halflife2 vs. halflife1	gauss1	1.8630e+03	0.1297	47	0.0927	6.2958	3			

Figure 11: Gaussian classification of Estimated half-life

4. CONCLUSIONS

Interested in the application of SVM in drug discoveries and designing we have examined two dataset of proteins with their various physic-chemical properties. Some binary classification problems do not have a simple hyper plane as a useful separating criterion. For those problems, there is a variant of the mathematical approach that retains nearly all the simplicity of an SVM separating hyper plane. The mathematical approach using kernels relies on the computational method of hyper planes. All the calculations for hyper plane classification use nothing more than dot products. Therefore, nonlinear kernels can use identical calculations and solution algorithms, and obtain classifiers that are nonlinear. The resulting classifiers are hyper surfaces in some space S. In our research from the sequences 100 different proteins we have generated two datasets which are then classified with their physico-chemical properties independently. We have generated one class of points inside the unit disk in two dimensions, and another class of points in the annulus from radius 1 to radius 2. Then a classifier is generated on the basis of the data with the Gaussian radial basis function kernel. The default linear classifier is obviously unsuitable for this problem, since the model is circularly symmetric. So we have set the box constraint parameter to Inf to make a strict classification, meaning no misclassified training points. Our results in the form of graphs have shown the non-linear classification of two type of data sets and then with use of curve fitting tool various other parameters are shown up which distinguish the two data sets accurately.

ACKNOWLEDGMENTS

I wish to express sincere gratitude and indebtedness to Supervisor, Prof. Rajbir Singh (Assoc. Prof. & Head, Department of Information Technology) for his valuable guidance, attention-grabbing views and obliging nature which led to the successful completion of this study. I lack words to express my cordial thanks to the members of Departmental Research Committee (DRC) for their useful comments and constructive suggestions during all the phases of the present study as well as critically going through the manuscript.

Words fail the author to express his deep sense of gratitude towards his family members for their moral and financial support and encouragement without which the author would not have been able to bring out this thesis.

REFERENCES

[1] B. Scholkopf et al. (1997) *"Comparing Support Vector Machines with Gaussian Kernels to Radial Basis Function Classifiers*", Signal Processing, IEEE Transactions Volume.45 Issue.11 pp.2758 - 2765.

[2] Gulsah Altun et al.(2006) *"Hybrid SVM kernels for protein secondary structure prediction "*, Granular Computing, 2006 IEEE International Conference pp. 762 - 765

[3] Y.Radhika et al.(2009)*"Atmospheric Temperature Prediction using Support Vector Machines"*, International Journal of Computer Theory and Engineering, Vol. 1, pp:793-820.

[4] Haibin Cheng et al(2010) *"Efficient Algorithm* for *Localized Support Vector Machine"* , Knowledge and Data Engineering, IEEE Transactions Volume:22 , Issue: 4 , pp: 537 - 549

[5] Li Liao (2003) *"Combining pairwise sequence similarity and support vector machines for remote protein homology detection"*, JOURNAL OF COMPUTATIONAL BIOLOGY Volume 10, pp: 857–868.

[6] Mehmet Gonen(2011) , *"Multiple Kernel Learning Algorithms*", Journal of Machine Learning Research, Vol.12, pp.2211-2268.

[7] Chih-Wei Hsu et al (2010), *"A Practical Guide to Support Vector Classification"* , National Taiwan University, Taipei 106, Taiwan,

[8] Asa Ben-Hur(2010) *"A User's Guide to Support Vector Machines"*, 10th ACM International conference of Computer and Information Technology (CIT) , pp.155- 162 .

[9] Christina Leslie et al.(2004) *"Mismatch string kernels for discriminative protein classification"*, Journal Bioinformatics Volume:20 Issue: 4, pp: 467-476

[10] Jason Weston (2004) *"Semi-supervised protein classification using cluster kernels*", Vol. 7, No.3, pp.1–8.

[11] Alexandros Karatzoglou(2006) *"Support Vector Machines in R"*, Journal of Statistical Software, Vol.15, Issue. 9.

[12] Lars Juhl Jensen et al. (2003) *"Functionality of System Component"* Conservation of Protein Function in Protein Feature Space Genome Res. 13(11): 2444–2449.

[13] Rong She et al. (2003) *"Frequent-Subsequence-Based Prediction of Outer Membrane Proteins"* Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 436-445.

[14] Ravi Gupta et al. (2008) *"A Time-Series-Based Feature Extraction Approach for Prediction of Protein Structural Class*" EURASIP J Bioinform Syst Biol 2008:235451.

[15] Kai-Yan Feng et al(2005) *"Boosting classifier for predicting protein domain structura lclass"* Biochemical and Biophysical Research Communications 334 pp:213–217

[16] Manpreet Singh et al. (2012) *"Machine Learning Classifiers for Human Protein*" International Journal of Computer Science and Telecommunications Volume 3, Issue 10 pp:21-25

[17] Andrey Fomenko et al.(2006)*"Prediction of protein functional specificity without an alignment*" OMICS Volume: 10 Issue:1 pp: 56-65

[18] Manpreet Singh et al.(2007) *"Human Protein Function Prediction using Decision Tree"* IJCSNS International Journal of Computer Science and Network Security, VOL.7 No.4 pp:92-98.

[19] Leslie et al. (2003) *"Mismatch string kernels for SVM protein classification" Neural Information Processing Systems 15*", pp. 1441-1448

[20] Ankita Srivastava et al.(2014) *"A Novel Data Mining Approach for Protein Function Prediction*" International Journal of Emerging Research in Management &Technology ISSN: 2278-9359 (Volume-3, Issue-5)

[21] Terrence S. Furey et al (2000) **"***Support vector machine classification and validation of cancer tissue samples using microarray expression data"* oxford journal vol:16, issue:10, pp:906-914.

[22] Datta S (2004) *"DFT based DNA splicing algorithms for prediction of protein coding regions"*, Page(s): 45 - 49 Vol.1, Conference Publications.

[23] Jennifer Harrow et al. (2009) *"Identifying protein-coding genes in genomic sequences"* Genome Biology.

[24] Mehmet Gonen et al. (2011) *"Multiple Kernel Learning Algorithms"* Journal of Machine Learning Research Vol:12 pp: 2211-2268

AUTHORS

Rajbir Singh is an Associate Professor & Head, Department of Information Technology of Lala Lajpat Rai Institute of Engineering & Technology Moga (cheema_patti@yahoo.com), India. He received his B.E (Honor) degree in Computer Science and Engineering from MD University, Rothak, Haryana and M-Tech degree in Computer Science and Engineering from Punjab Technical University, Jalandhar Pb. (INDIA). He has authored 03 books on Computer Science. His main field of research interest is Bio-Informatics and Data mining. He works on the Gene Expression, Phylogenetic Trees and Prediction of Protein Sequence & Structure.

Ravdeep Singh is student of M.Tech Department of Computer Science & Engg. Lala Lajpat Rai institute of engg. & Tech, Moga (ravdeep2323@yahoo.com), Punjab, INDIA. He received his B.Tech in Computer Science & Engineering degree from Punjab Technical University, Jalandhar Pb. (INDIA). His research interest includes Bio-Informatics, Software Engineering, & Software Testing. He works on the Improved Protein Function Classification Using Support Vector Machine.