Performance Investigation of Support Vector Regression using Meteorological Data

Somya Jain[1] and MPS Bhatia[2]

[1]Division of Computer Engineering
Netaji Subas Institute of Technology, New Delhi
University of Delhi
[2]Division of Computer Engineering
Netaji Subas Institute of Technology,
New Delhi
University of Delhi
[1]Somya.nsit@gmail.com, [2]Mps.bhatia@gmail.com

Abstract

Predicting fire nature is artistry as much as it's a science. Forecasting the burnt area and range of field plays a vital role in resource abatement and renewal efforts. Literature studies have shown that machine learning techniques achieved better performance in forecasting and trend perusal. The purpose of this paper is to investigate the relevance of the state-of-the-art machine learning techniques epsilon Support Vector Regression and Nu-SVR to predict forest fire occurrence and burned area utilizing the meteorological data. The goals of this research are to (1) Identifying the best parameter settings using a grid-search and pattern search technique; (2) comparing the prediction accuracy among the models using different data sorting methods, random sampling and cross-validation. In conclusion, the experiments show that E-SVR performs better using various fitness-functions and variance analysis. The study is carried out to build predictive models for guesstimating the risk of the outbreaks in Montesinho Natural Park.

Keywords: *Forest Fire Prediction, Machine Learning, Regression, Support Vector Machine,* Epsilon SVR, Nu-SVR

1. Introduction

Forest fires are one of the most important and prevalent type of disasters and they can create great environmental problems for Nature. In this work Support Vector Machines (SVM) which has gained profound interest among the researchers is intended to predict forest fire burnt area by deducting the number of weather parameters utilizing geospatial data with the highest prediction accuracy. [1] Support Vector Machines (SVMs) are a set of related supervised learning methods. SVM was first heard in 1992, introduced by Vladimir Vapnik [2] and colleagues Bernhard Boser and Isabelle Guyon in the first paper on SVM presented by them but the ground works for SVMs has been around since the 1960s (including the work by Vapnik and Alexei Chervonekis on Statistical Learning Theory). [3] SVM is a classification and regression prediction tool that uses machine learning theory to amplify predictive precision while automatically abstain over-fitting problem of data. SVM is now a bustling part of machine learning research around the world. SVM can be defined as the system which uses a nonlinear mapping to transform the original training data into higher dimension. It searches for the linear optimal separating hyper plane (that is, a "decision boundary separating the tuples of one class from another") within this new higher dimension.

The SVM finds this hyper plane using support vectors ("essential" training tuples) and margins (defined by the support vectors) [4]. Some of the applications of SVM include handwritten digit identification, speaker recognition, object identification and benchmark time-series prediction tests. SVM has the better ability to generalize which is the goal in demographic learning. Results of some experiments are also represented.

2. Literature Review

Through these various issues and challenges in forest fire occurrence prediction are highlighted which gives an overview of existing studies related to these measures and the whole section throws light on some machine learning techniques. Forest fire databases started being developed years ago and their main target was to record forest fires, burnt area and associated weather parameters. The proposed work has been motivated by several earlier researches in the literature related to forest fire detection using spatial data and machine learning techniques. [5] Cellular automata and fractal geometry are applied by Clarke *et al.*, (1994) to predict wild fire propagation and extinction. [6] Neural Networks (NN) and logistic regression models are applied by Vega-Garcia *et al.*, (1996) to predict human-caused wildfire occurrence in Canada with the accuracy of 76% of the fire and non-fire observations on the test data and to identify the important input variables. [7] Wiering and Dorigo (1998) used neural networks and parameters such as fuel-type, wind speed to build the spread index and to minimize the damage done by the forest fire by knowing where to cut fire-line. [8] Felber and Bartelt (2003) used the *k*-Nearest Neighbor algorithm to compare past fire occurrences to current forecast conditions in order to predict forest fire danger. [9] Cortez and Morais (2007) predicted the burned area of forest fires using SVM and Random Forests. Four distinct feature selection setups were tested on recent real-world data collected from the northeast region of Portugal. Their configuration uses four meteorological inputs (*i.e.*, temperature, relative humidity, rain and wind) and is capable of predicting the burnt area of small fires which are more frequent. [10] Cheng and Wang (2008) presented an integrated spatio-temporal forecasting framework that uses dynamic recurrent neural networks for forecasting the annual average area of forest fire utilizing the historical data. [11] Daniela Stojanova predicts forest fires in Slovenia using different data mining techniques. The authors have employed the predictive models based on the data from a GIS (Geographical Information System) and the weather prediction model. The work examined three different datasets including Kras region, Primorska region and for continental Slovenia. The researchers demonstrated that Bagging and boosting of decision trees offers the best results in terms of accuracy for all three datasets. [12] Markuzon and Kolitz (2009) used Random Forests, Bayesian networks and the k-Nearest Neighbor method for estimating fire danger. They used data from images and weather information. None of the classifiers showed significantly superior performance over the others. However, the study demonstrated a significant predictive power of fire models that are based on remote sensing observations. [13] George E. Sakr (2011) applied ANN and SVM for the forest fire occurrence prediction using two weather parameters and shows that ANN outperforms SVM on average by 0.17 fires, while SVM outperforms ANN in the binary classification.

3. Predictive Model

In this, we will examine the Support Vector Regression (SVR) as a predictive model for detecting and predicting the forest fire burnt area with large data set. As prediction is an intricate and challenging task for researchers and the same is extremely important in this forest fire context as the expense of the misclassification using a classifier is sky high. Support Vector Regression (SVR), a category for support vector machine can be applied to regression problems with the introduction of an alternative loss function including a distance measure and attempts to minimize the error bound so as to achieve best performance [14] [15]. On the basis of training sample regression is finding a function which approximates mapping from an input domain to the real numbers. Support vector regression is the extended version of large margin kernel methods used for regression analysis. [15] The regression can be classified into linear and non linear regression. Linear models mainly consist of the e-intensive, Huber and quadratic loss function. In the same manner of the non-linear SVR approach, a non-linear mapping can be used to map the data into a high dimensional feature space where linear regression is performed. To address the curse of dimensionality kernel approach is employed. Regression methods are based on prior knowledge of the problem and the distribution of the noise. Instead of attempting to classify new unseen variables $x^{'}$ into one of two categories $y^{'} = \pm 1$, we now wish to predict a real valued output for $y^{'}$ so that our training data is of the form [12]:

$$D = \{(x_i, y_i), ..., (x_l, y_l)\}, x \in R^n, y \in R \quad (1)$$

$$\{x_i, y_i\}, \text{ where } \ i = 1.....l$$

With a linear function:- $y_i = W.x_i + b$

The optimal error regression function is given by the minimum of the function,

$$\varphi(W, x) = \frac{1}{2}\|W\|^2 + C\sum_i (\varepsilon_i^- + \varepsilon_i^+) \quad (2)$$

Where C is a pre-specified value, and $\varepsilon^-, \varepsilon^+$ are slack variables out of which one is needed to control the error induced by the observations that are larger than the upper bound of the ε-tube and the other for the observations that are smaller than the lower bound. There are two types of SVM regression epsilon-SVR and Nu-SVR. In Epsilon-SVR ε parameter is used to employ a penalty to the accession for points which were not accurately predicted [16] where as in Nu-SVR the ε parameter was replaced by an alternative nu parameter, which utilize different penalty to control the number of support vectors but the decision function is same as that of ε-SVR. Both can be used for density estimation and similar accretion problem is solved in both the cases.

3.1. Data Set Resource

In this work forest fire data from the Montesinho Natural Park Tras-os-Montes north east region of Portugal will be considered [9 ,17]. The data used for this study comprises the data collected from January 2000 to December 2003 and it was built using two sources .The first database on the fire occurrence were provided by the inspector that was responsible for Montesinho fire occurrence. On a daily basis, every time a forest fire occurred, several features such as time, date, spatial location, the type of vegetation involved, the six components of FWI System and the total burned areas were registered. The second database was collected by the Braganca Polytechnic Institute, which were aggregated into a single dataset with a total of 517 entries based on the several weather observations. Environmental data describe the environment of the outbreaks including Geographical (GIS), Temporal, FWI, Meteorological data. The spatial unit of the analysis was within a 9×9 Grid and all the attributes were aggregated to this resolution. Meteorological data consists of four weather attributes (temperature, RH, wind, rain) from the meteorological station database.

Table 1. Preprocessed Dataset Attributes

Attribute	Range
X and Y-axis coordinate	1 to 9
Month-month of the year	January to December
Day-day of the week	Monday to Sunday
FFMC code	18.7 to 96.2
DMC code	1.1 to 291.3
DC code	7.9 to 860.6
ISI index	0 to 56.1
Temperature (degree centigrade)	2.2 to 33.30
Relative humidity (percentage)	15.0 to 100
Wind speed (km/hr)	0.40 to 9.40
Rain (mm/m2)	0.0 to 6.4
Burned area size (ha)	0.00 to 1090.84

4. Experiments and Results

All the experiments reported in this study were conducted using the DTREG, a Predictive Modeling Software which builds classification and regression models that describe data relationships and can be used to predict values for future observations [18]. It facilitates the use of data mining techniques. Before fitting the model, some preprocessing was required by the SVM model. Next the support Vector Regression model was fitted. Particularly in this work of E-SVR and Nu-SVR models, a grid search and pattern search techniques are used for finding optimal parameter values. A grid search seeks values of each parameter using geometric steps across the search range. Model parameters such as upper bound C with the search range 0.1 to 5000, kernel parameter γ to the 0.01 to 50, P (Epsilon) with 6.001 to 10^2 and polynomial degree 1 for analysis and Nu having the range 0.001 to 0.6. [18] A pattern search starts at the center of the search range and makes trial steps in each direction for each parameter. The search center moves to the new point and the process is iterated if the fit of the model improves. If there is no improvement the step size is reduced and the search is tried again. When the search step size is reduced to the specified tolerance the pattern search stops. Selecting the pattern search technique using 10 search intervals which would require 1000 model evaluations and 1e-008 tolerance for stopping the iterative optimization process and the optimal values through the grid search we could build a forest fire prediction model with the higher stability and prediction power. Model parameters can significantly impact the accuracy of the model. If cross-validation is used for each model evaluation, the number of actual calculations would be further multiplied by the number of cross-validation folds (10 for our experiment setup). Missing values are handled by replacing missing values with medians. 100% rows are used for searching and goal is to minimize the total error. For model testing and validation 10 cross-validation and random percent hold-back (20%) is used. Other miscellaneous controls used were (1) shrinking heuristics which improve performance when the training data is large by ignoring the points that are far from overlapping and which are unlikely to influence the choice of the optimal separating hyperplane. Essentially, shrinking eliminates outlying vectors from consideration. (2) Cache with default value of 256 MB to store truncated rows of the recorded kernel matrix. This cache avoids recomputing components of the kernel matrix and can speed up the computation by a significant amount in some cases. Variable importance is calculated by generating the report on the relative significance of predictor variables (Figure 1). After calculating the variable importance for predictor variable it is better to use variables (temp, RH, wind, and rain) rather than FWI Variables. The combination of this gives highest prediction accuracy. The error for each prediction is the difference between the actual area and the predicted area (Figure 2).

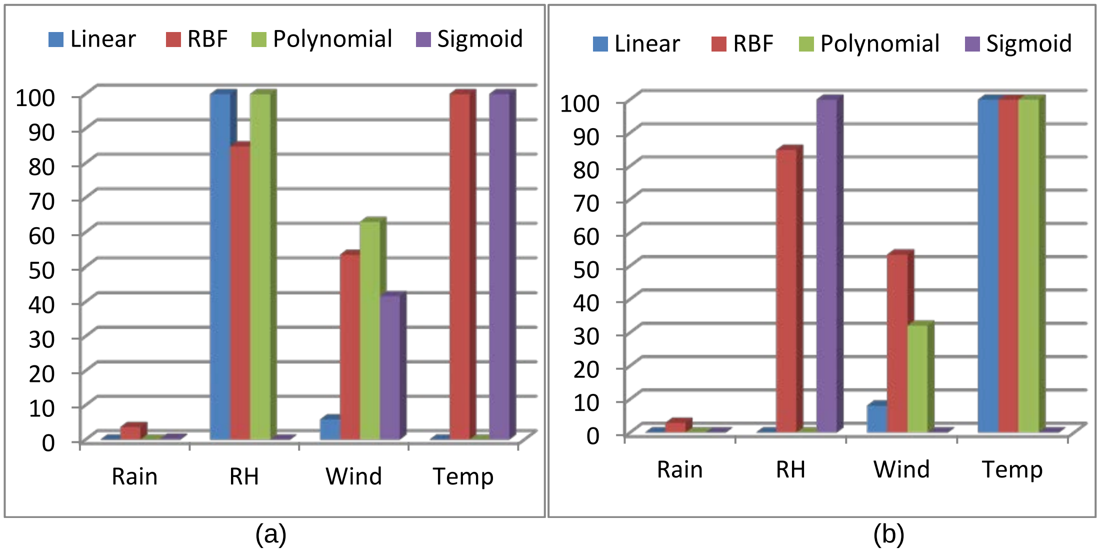


Figure 1. Relative Importance of Predictor Variables for (a) E-SVR and (b) Nu-SV

The fitness value is an indicator of error; ∴ smaller values indicate higher fitness. We calculate the RMSE, MAE, MSE, NMSE and MAPE as the measure of fitness. MAPE is the more objective statistic indicator because the measure is in relative percentage and will not be affected by the unit of the forecasting series. To validate the performance of the models, we statistically compared its prediction accuracy, analysis of variance, fitness functions with 10-fold cross-validation (Table 2) and random sampling (Table 3). Table 4 compares the averaged results after running 10-fold cross validation with RBF kernel to the Results published by Cortez and Morais in previous work on this data set [9]. This comparison demonstrates that our predictive model produces better results than the methodologies used in previous work when RMSE is used as the measure of error, but does not perform as well when MAE is used as a measure (Figure 3).

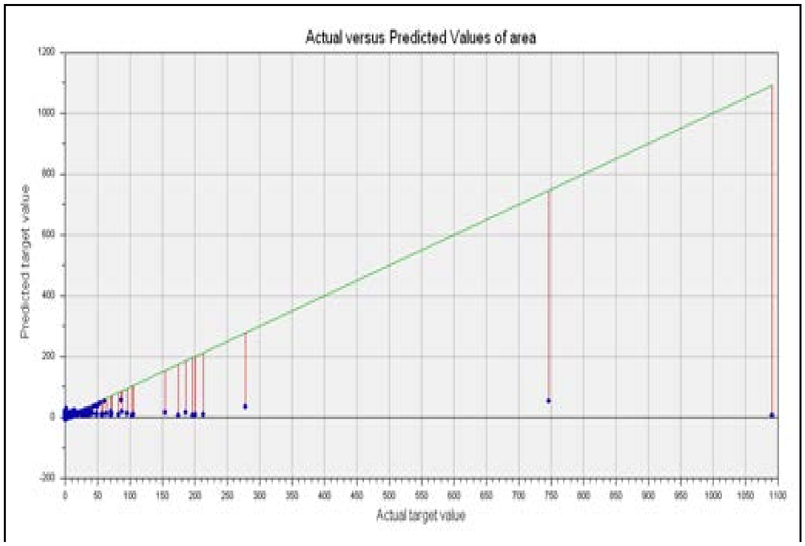


Figure 2. Actual versus Predicted Values of the area using E-SVR

Table 2. Comparison on the Basis of Fitness Function with 10-fold Cross Validation

10 cross validation folds				
Fitness Function	E-SVR			
	Linear Kernel	RBF Kernel	Polynomial Kernel	Sigmoid Kernel
RMSE	63.5942	63.5144	63.5942	63.5416
MSE	4044.2252	4034.0791	4044.2248	4037.5414
MAE	18.5668	16.2207	18.5669	19.1787
MAPE	505.9584	365.8025	505.9616	609.7511
NMSE	1.0000	0.997491	1.0000	0.9983
Fitness Function	Nu-SVR			
	Linear Kernel	RBF Kernel	Polynomial Kernel	Sigmoid Kernel
RMSE	63.6090	63.4304	63.6045	63.5485
MSE	4046.107	4023.4273	4045.5412	4038.4122
MAE	18.6270	16.3133	18.5940	18.3254
MAPE	509.8537	376.3455	508.6723	500.0750
NMSE	1.0004	0.9948	1.0003	0.9985

Table 3. Comparison on the Basis of Fitness Function with Random Sampling (20%)

Random Sampling (20%)				
Fitness Function	E-SVR			
	Linear Kernel	RBF Kernel	Polynomial Kernel	Sigmoid Kernel
RMSE	109.1259	110.4509	109.1259	108.4329
MSE	11908.468	12199.416	11908.466	11929.879
MAE	23.7947	21.7922	23.7945	24.4573
MAPE	449.4660	352.8658	449.4544	378.0400
NMSE	1.0053	1.0298	1.0053	0.9926
Fitness Function	Nu-SVR			
	Linear Kernel	RBF Kernel	Polynomial Kernel	Sigmoid Kernel
RMSE	109.2239	110.4036	109.2247	109.0895
MSE	11929.879	12188.956	11930.037	11900.527
MAE	22.9800	21.8763	22.9569	24.0549
MAPE	395.9012	374.2929	394.4423	481.3810
NMSE	1.0071	1.0290	1.0071	1.0046

Table 4. Comparison of Results with Previous Work

Measure of Fitness	Cortez and Morais [9]	Our Algorithm	
		E-SVR	Nu-SVR
RMSE	64.7	63.5	63.4
MAE	12.71	16.22	16.31

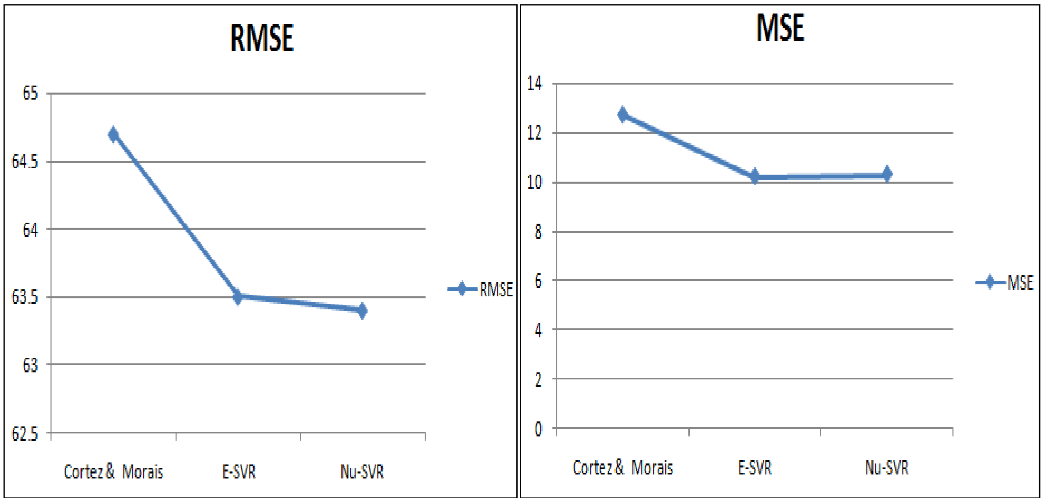


Figure 3. RMSE (left) and MSE (right) Comparison Graph

The time complexity of different kernel functions with Cross-validation and Random Sampling method are also compared among the models. With these results, E-SVR can serve as a promising alternative for the forest fire occurrence prediction except with the RBF Kernel for which Nu-SVR performs better. The analysis of time complexity of the models is presented in (Figure 4). It should be noted that appropriate kernel function [19] can be problem specific; while during this work it is observed that the polynomial kernel function takes a longer analysis run time and linear kernel function takes the shortest running time. In addition, we demonstrate the accuracy of other kernel functions such as linear, polynomial, (Radial basis function) RBF and sigmoid functions to validate the RBF kernel function as adoption. A cross-validation result outperforms the random sampling results.

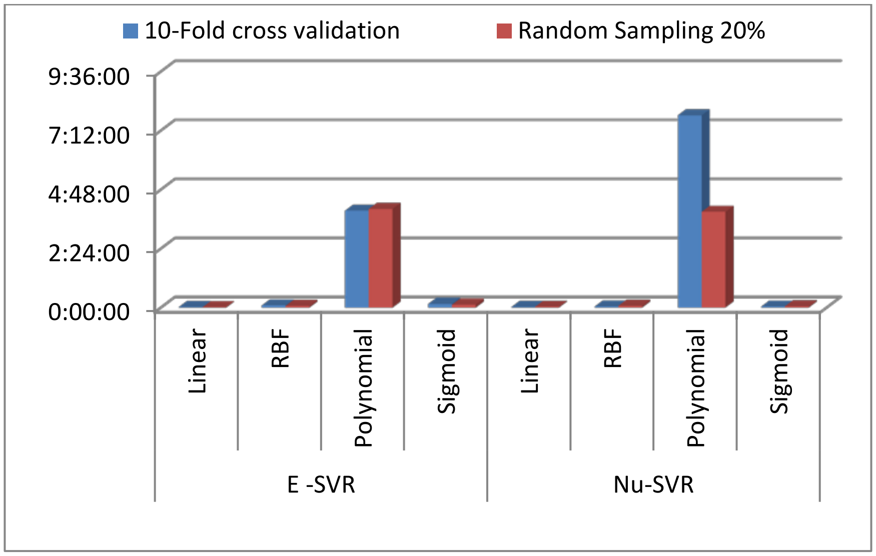


Figure 4. Comparison of Time Complexity with Different Kernel Functions

5. Conclusion

This paper presented an overview of SVR which provide a new approach to the problem of forest fire burnt area prediction (together with regression Estimation) with a clear association to the underlying statistical learning theory. The results demonstrated that E-SVR and Nu-SVR differs radically from comparable approaches: SVR training and their geometric perception provides fertile ground for further investigation. An SVR is largely characterized by an eclectic of its kernel. The kernel mapping provides the framework for most of the commonly employed models, enabling comparisons to be performed. Future scope includes a method for selecting the kernel function, capacity checker and advancement of kernels with invariance's.

References

[1] V. Jakkula, "Tutorial on Support Vector Machine (SVM)", School of EECS, Washington State University, Pullman, 99164.

[2] T. Fletcher, "Support Vector Machines", UCL www.cs.ucl.ac.uk/sta_/T.Fletcher/.

[3] J. Han and M. Kamber, "Data Mining Concepts and Techniques".

[4] C. Cortes and V. Vapnik, "Support Vector Networks", AT& T labs-research, USA.

[5] K. Clarke and R. J. Brass, "A cellular automaton model of wildfire propagation and extinction", Photogrammetric Engineering and Remote Sensing, **(1994)**, pp. 1355-1367.

[6] C. Vega-Garcia, B. Lee and T. P. Woodard, "Applying neural network technology to human-caused wildfire occurrence prediction", AI Appl., vol. 10, no. 3, **(1996)**, pp. 9-18.

[7] M. Wiering and M. Dorigo, "Learning to control forest fires", Proceedings of the 12th International Symposium on Computer Science for Environmental Protection, **(1998)**, pp. 378-388.

[8] B. A. Felber, "The use of nearest neighbor method to predict forest fires", Proceedings of the 4th international workshop on remote sensing and GIS applications to forest fire management: innovative concepts and methods in fire danger estimation, **(2003)**, pp. 100-10.

[9] M. P. Cortez, "A data mining approach to predict forest fires using meteorological data", **(2007)**.

[10] W. T. Cheng, "Integrated spatio-temporal data mining for forest fire prediction", Trans GIS, vol. 12, no. 5, **(2008)**, pp. 591-611.

[11] D. Stojanova, A. Kobler and P. Ogrinc Bernard, "Estimating the risk of fire outbreaks in the natural environment", Springers.

[12] N. Markuzon and S. Kolitz, "Data driven approach to estimating fire danger from satellite images and weather information", 38th IEEE applied imagery pattern recognition workshop, **(2009)**.

[13] G. E. Sakra, I. H. Elhajj and G. Mitri, "Efficient forest fire occurrence prediction for developing countries using two weather parameters", Engineering Applications of Artificial Intelligence, ElsevierLtd, vol. 24, **(2011)**, pp. 888-894.

[14] C.-W. Hsu, C.-C. Chang and C.-J. Lin, "A Practical Guide to Support Vector Classification", Taipei 106.

[15] S. R. Gunn, "Support Vector Machines for Classification and Regression", University of Southampton Technical Report.

[16] M. Pirooznia, "Expression sequence tags analysis, annotation and toxic genomics and learning".

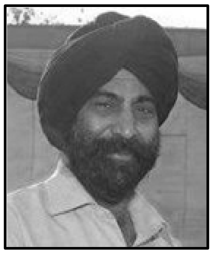
[17] UCI machine learning repository www.ics.uci.edu/~mlearn/.

[18] P. H. Sherrod, "DTREG Predictive Modeling Software", Copyright © 2003-2013, www.dtreg.com.

[19] A. Ng, Lecture notes on Support Vector Machines, Part V. CS229.

Authors

Somya Jain received the B.Tech degree in 2006 from Kurukshetra University, Kurukshetra. She was a research student of Netaji Subas Institute of Technology, University of Delhi, India in 2013. She is currently an Assistant Professor at the department of Computer Science, Jaypee Institute of Information Technology, Noida, India. Her research interest is in the area of Machine Learning, Software Engineering, data and text mining.



M.P.S Bhatia is currently Professor at the department of Computer Science, Netaji Subas Institute of Technology, University of Delhi, India. He has decades of experience in the field of Computer Science and Information Technology. He is specialized in setting up R&D and innovation. He has supervised numerous PhD and M.Tech research fellows. He is having a large number of publications. His research area is Machine Learning, Information retrieval and text mining.