Support Vector Regression to Forecast the Demand and Supply of Pulpwood

V. Anandhi and R. Manicka Chezian

***Abstract*—Support Vector Machine (SVM) is a popular machine learning method for classification, regression, and other learning tasks. Support Vector Regression (SVR), a category for support vector machine attempts to minimize the generalization error bound so as to achieve generalized performance. Regression is that of finding a function which approximates mapping from an input domain to the real numbers on the basis of a training sample. Support vector regression is the natural extension of large margin kernel methods used for classification to regression analysis. In this paper Support Vector Regression is used to forecast the demand and supply of pulpwood. The usage of paper increases day to day. Wood Pulp is the most common raw material in paper making. On account of steady increase in paper demand, the forecast on demand and supply of pulp wood is considered to improve the socio economic development of India. Forecasting is done in Libsvm a library for support vector machines by integrating it with MATLAB.**

***Index Terms*—Support vector machines (SVM), support vector regression (SVR), wood pulp, forecast, kernel.**

I. Introduction

Support Vector Machines (SVM) is learning machines implementing the structural risk minimization inductive principle to obtain good generalization on a limited number of learning patterns. The theory has originally been developed by Vapnik[1] and his co-workers on a basis of a separable bipartition problem at the AT & T Bell Laboratories. SVM implements a learning algorithm, useful for recognizing subtle patterns in complex data sets. Instead of minimizing the observed training error, Support Vector Regression (SVR) attempts to minimize the generalization error bound so as to achieve generalized performance. There are two main categories for support vector machines: support vector classification (SVC) and support vector regression (SVR). SVM is a learning system using a high dimensional feature space. It yields prediction functions that are expanded on a subset of support vectors. SVM can generalize complicated gray level structures with only a very few support vectors and thus provides a new mechanism for image compression. A version of a SVM for regression has been proposed in 1997 by Vapnik, Steven Golowich, and Alex Smola [2]. This method is called support vector regression (SVR)the model produced by SVR only depends on a subset of the training data, because the cost function for building the model ignores any training data that is close (within a threshold ε) to the model prediction [3]. Support Vector Regression (SVR) is the most common application form of SVMs. Support vector machines project the data into a higher dimensional space and maximize the margins between classes or minimize the error margin for regression [4].

Manuscript received September 4, 2012; revised November 18, 2012.

V. Anandhi is with the Department of Forest Resource Management, Forest College and Research Institute, Mettupalayam 641 301, Tamil Nadu, India (e-mail: anandhivenugopal@gmail.com).

R. Manicka Chezian is with the Department of Computer Science, NGM College, Pollachi- 642 001, Tamil Nadu, India.

II. Literature Review

Support Vector Machines (SVMs) are a popular machine learning method for classification, regression, and other learning tasks. Basic principle of SVM is that given a set of points which need to be classified into two classes, find a separating hyperplane which maximises the margin between the two classes. This will ensure the better classification of the unseen points, i.e. better generalisation..In SVR, our goal is to find a function f(x) that has at most deviation from the actually obtained targets yi for all the training data. Forecasting is a systematic effort to anticipate the future events or conditions. Forecasting involves rainfall forecasting, stock market – price forecasting, temperature forecasting, cash forecasting in bank etc. Forecasting is usually carried out using various statistical analysis. In this paper SVR is used for forecasting the demand and the supply of pulpwood and the forecasting is done using libsvm. Libsvm is a library for support vector machines. LIBSVM is currently one of the most widely used SVM software.

In India there are about 600 paper mills and out of which 30 to 40% of the industries use wood[5] as a raw material predominantly. The current raw material recruitment is more than 5 million metric cube against the domestic supply of 2.6 million metric cube created a short fall of more than 45%. The short fall is met mostly from imports. Similarly in Tamil Nadu, there are about 39 paper mills and of which only 2 paper mills[6] are wood based and the demand is around 8 lakhs tones of wood against the domestic supply of less than 4 lakhs tones of wood. The demand is increasing at an alarming rate without increase in the actual supply. Three main pulp wood based industries in Tamil Nadu are Tamilnadu News print and Papers Limited (TNPL), karur District, Seshasayee paper and Boards (SPB), Erode District, South India Viscose industries Limited, Mettupalayam, Coimbatore District.Paper and paper board industries are classified into two groups[7] namely, cultural paper used for writing purpose and Industrial paper for wrapping, packaging purposes. Newsprint is a low-cost, commonly used to print newspapers, and other publications and advertising material. The demand for pulp wood is mainly from the pulp, paper and the newsprint industry. The major supplies of pulp wood firms in Tamil Nadu are State Forest Department, Tamil

Nadu State Forest Plantation Corporation (TAFCORN), Private Plantations.

III. Methodology

Support vector regression is the natural extension of large margin kernel methods [8] used for classification to regression analysis. The problem of regression is that of finding a function which approximates mapping from an input domain to the real numbers on the basis of a training sample. This refers to the difference between the hypothesis output and it's training [9] value as the residual of the output, an indication of the accuracy of the fit at this point. One must decide how to measure the importance of this accuracy, as small residuals may be inevitable even while we need to avoid in large ones. The loss function determines this measure. Each choice of loss function will result in a different overall strategy for performing regression.

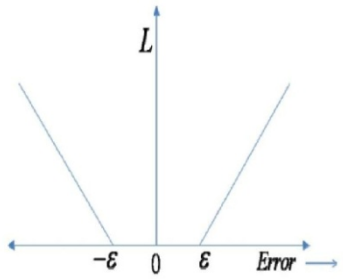

Fig. 1. ε -insensitive Loss Function for regression.

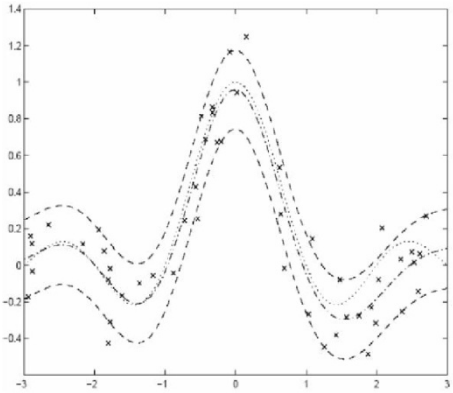

Fig. 2. ε -insensitive zone for non-linear support vector regression.

Support vector regression performs linear regression in the feature space using ε - insensitive loss function and, at the same time, tries to reduce model complexity by minimising || w ||2. This can be described by introducing (non-negative) slack variables ξ_i, ξ_i^* i=1,…n to measure the deviation of training samples outside the ε – insensitive zone.

The SV regression is formulated as

$$Min \frac{1}{2}|| w ||^2 + C \sum_{i=1}^{n} (\xi_i + \xi_i^*)$$

Subject to

$$w^T \emptyset(x_i) - y_i \leq \varepsilon + \xi_i$$
$$y_i - w^T \emptyset(x_i) \leq \varepsilon + \xi_i^*$$

$$\xi_i, \xi_i^* \geq 0, \text{ i= 1,....m}$$

Φ maps the data point x into high-dimensional feature space and linear regression with ε - insensitive loss function) is performed in that feature space. As a consequence, the dimension of w is that of φ(w) and hence there is no attempt to express the target function in terms of w.

TABLE I: Sample Normalized data of the Demand and Supply of Pulp Wood in MT (Metric Tonnes)

Year	Demand	Supply
0.9945	0.2778	0.2716
0.9950	0.2889	0.3221
0.9955	0.2972	0.2790
0.9960	0.3278	0.2734
0.9965	0.3662	0.3621
0.9970	0.3699	0.4670
0.9975	0.4013	0.4944
0.9980	0.8518	0.7099
0.9985	0.8749	0.8782
0.9990	0.9414	0.9247

MATLAB is a high-level language and provides an interactive environment that enables us to perform computationally intensive tasks faster than with traditional programming languages.A MATLAB interface for SVM is Libsvm [10]. Libsvm-mat-2.88-1 is integrated with Matlab. The data for forecasting are to be collected. The data need to be split the data set into two, one for training and other for testing. After collecting the data, we need to convert both the training set and testing set into SVM format. The SVM algorithm operates on numeric attributes. So we need to convert the data into libsvm format which contains only numerical values. The original data maybe too huge or small in range, thus we can rescale them to the proper range so that training and predicting will be faster. The main advantage of scaling is to avoid attributes in greater numeric ranges dominating those in smaller numeric ranges. After scaling the data set, we have to choose a kernel function for creating the model. The basic kernels are linear, polynomial, radial basis function, sigmoid. In general, the RBF kernel is a reasonable first choice. A recent result shows that if RBF is used with model selection, then there is no need to consider the linear kernel. The kernel matrix using sigmoid may not be positive definite and in general it's accuracy is not better than RBF. Polynomial kernels are ok but if a high degree is used, numerical difficulties tend to happen. A typical use of LIBSVM involves two steps: first, training a data set to obtain a model and second, using the model to predict information of a testing data set. SVM train trains a support vector machine. Set the SVM type to epsilon SVR. There are two commonly used versions of SVM regression, 'epsilon-SVR' and 'nu-SVR'. The original SVM formulations for Regression (SVR) used parameters to apply a penalty to the optimization for points which were not correctly predicted. An alternative version of both SVM regression was later developed where the epsilon penalty parameter

was replaced by an alternative parameter, which applies a slightly different penalty. Epsilon or nu are just different versions of the penalty parameter.

IV. Results and Discussion

The study is based on the data collected at the Tamil Nadu News print and papers Limited (TNPL) in Karur District, Tamil Nadu. Since the demand for paper increases rapidly and woodpulp is largely used for making paper, data of the demand and supply of pulpwood were collected for over fifteen years for forecasting. The use of Support Vector regressions for forecasting is increasing rapidly. The libsvm is integrated with MATLAB and the forecasting is done with it.

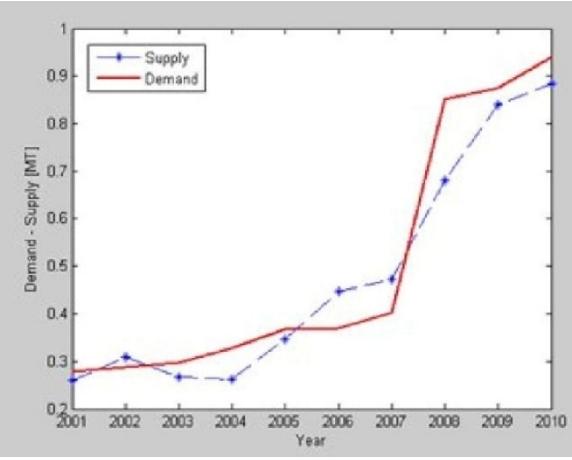

Fig. 3. Demand and supply patterns for the ten years data.

The demand and the supply patterns collected were normalized. The support vector train function was used to train the model. In order to get an optimized value of output, the tuning parameters are gamma, cost, kernel type (RBF kernel), Degree of kernel, Epsilon value.The year, demand and the supply patterns are plotted.

There were four types of kernels, linear, polynomial, radial basis function (RBF) and sigmoid, used for SVM models. Among these kernels, RBF was the most frequently used kernel because of their localized and finite responses across the entire range of the real x-axis [11].The type of the kernel function used is the radial basis function. Normally a Gaussian will be used as the RBF, it shows a two-dimensional version of such a kernel. The output of a kernel is dependent on the Euclidean distance of x_j from x_i. The support vector will be the centre of the RBF and σ will determine the area of influence this support vector has over the data space. A larger value of will give a smoother decision surface and more regular decision boundary. This is because an RBF with large σ will allow a support vector to have a strong influence over a larger area. The decision surface and boundaries for two different σ values. A larger σ value also increases the α value (the Lagrange multiplier) for the classifier. When one support vector influences a larger area, all other support vectors in the area will increase in α - value to counter this influence. Hence all α α- values will reach a balance at a larger magnitude. A larger σ –value α will also reduce the number of support vector. Since each support vector can cover a larger space, fewer are needed to define a boundary. Set the parameter C of C-SVC, epsilon-SVR, and nu-SVR (default 1). Increasing cost value causes closer fitting to the calibration/training data. Kernel *gamma* parameter controls the shape of the separating hyperplane. Increasing gamma usually increases number of support vectors.

In training the regress which are predicted within distance epsilon from the actual value. Decreasing epsilon forcession function there is no penalty associated with point closer fitting to the calibration/training data. The demand is forecasted to be 5, 00, 000 MT and supply 4, 67, 000 for the forthcoming year for TNPL. The value is then forecasted using the test data and the model created using the train data.

V. Conclusion

The study was conducted with the overall objective of the analysis and forecast of demand and the supply of pulpwood, so forest based industries can raise their raw materials to meet their needs through agroforestry programmes. A Support Vector Regression based prediction model appropriately tuned can outperform other more complex models. The study could further be extended to forecasting the demand and the supply of various species of wood and also their price. The awareness of the demand and supply patterns are a supportive mechanism which demanded a systematic forecasting system similar to agricultural products. Support vector regression is a statistical method for creating regression functions of arbitrary type from a set of training data. Testing the quality of regression on the training set has shown good prediction accuracy.

References

[1] V. Vapnik, *The nature of statistical learning theory*, Springer, NY, 2000.
[2] V. Vapnik, S. Golowich, and A. Smola, "Support vector method for function approximation, regression estimation, and signal processing," *Neural Information Processing Systems*, vol. 9, MIT Press, Cambridge, MA. 1997.
[3] D. Basak, S. Pal, and D. C. Patranabis, "Neural Information Processing," *Letters and Reviews*, vol. 11, no. 10, pp. 203-224, October 2007.
[4] "A Comparison of Machine Learning Techniques and Traditional Methods," *Journal of Applied Sciences*, vol. 9, pp. 521-527.
[5] *Forest Survey of India*, Ministry of Environment and Forests, Govt. of India, 2009.
[6] K. T. Parthiban and R. M. Govinda, "Pulp wood based Industrial Agroforestry in Tamil Nadu – Case Study, Indian Forester," 2008.
[7] V. Anandhi, R. M. Chezian, and K. T. Parthiban, "Forecast of demand and supply of pulpwood using artificial neural network," *International Journal of Computer Science and Telecommunications*, vol. 3, no. 6, pp. 35-38, 2012.
[8] K. P. Soman, R. Loganathan, and V. Ajay, "Support vector machines and other kernel methods," Centre for Excellence in Computational Engineering and Networking Amrita Vishwa Vidyapeetham.
[9] H. Drucker, C. J. C. Burges, L. Kaufman, A. Smola, and V. Vapnik, "Support vector regression machines," *Advances in Neural Information Processing Systems*, The MIT Press, vol. 9, pp. 155, 1997.
[10] C. C. Chang and C. J. Lin, *A library for support vector machines*, Technical Report, Department of Computer Science and Information Engineering, National Taiwan University, 2001.
[11] Z. Hua and B. Zhang, "A hybrid support vector machines and logistic regression approach for forecasting intermittent demand of spare parts," *Appl Math Comput*, vol. 181, pp. 1035-1048, 2006

R. Manickachezian received his M.Sc Applied Science from PSG College of Technology, Coimbatore, India in 1987. He completed his M.S. degree in Software Systems from Birla Institute of Technology and Science, Pilani, Rajasthan, India and Ph.D degree in Computer Science from School of Computer Science and Engineering, Bharathiar University, Coimbatore. He has 25 years of Teaching experience and 17 years of Research Experience. He served as a Faculty of Maths and Computer Applications at P.S.G College of Technology, Coimbatore from 1987 to 1989. Presently, he is working as an Associate Professor of Computer Science in NGM College (Autonomous), Pollachi, India. He has published 55 papers in various International Journals and Conferences. He is a recipient of many awards like Desha Mithra Award and Best paper Award . He is a member of various Professional Bodies like Computer Society of India and Indian Science Congress Association. His research focuses on Network Databases, Data Mining, Distributed Computing, Data Compression, Mobile Computing and Real Time Systems

V. Anandhi received M.C.A from R.V.S College of Engineering and Technology, Dindigul, India in 2001, M.Phil from Kongunadu Arts and Science college, Coimbatore, India in 2005. Presently working as Assistant Professor(Computer Science) at Tamil Nadu Agriculture University, Coimbatore, India. She has published papers in various International Journals and Conferences. Her Research focuses on Artificial Neural Networks, Databases, Data Mining and Support Vector Machines.