ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 10, October 2013

SUPPORT VECTOR REGRESSION IN FORECASTING

V. Anandhi[1] and Dr. R. Manicka Chezian[2]

Assistant Professor(Computer Science), Department of Forest Resource Management, Forest College and Research Institute, Tamil Nadu Agricultural University, Mettupalayam 641 301,Tamil Nadu, India[1]

Associate Professor, Department of Computer Science, NGM College, Pollachi-642001, Tamil Nadu, India[2]

Abstract: Support Vector Regression (SVR), a category for Support Vector Machine (SVM) 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. 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.

Keywords: Support Vector Regression (SVR), Support Vector Machine (SVM), regression, kernel, pulp wood

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. 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].

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 generalisationSupport vector machines (SVM) are used as they reduce the time and expertise needed to construct/train price forecasting models. Also SVM has lower tune-able parameters with parameter values choice being less critical for good forecasting results. SVM can optimize its structure (tune its parameter settings) on input training data provided. SVM training includes solving quadratic optimization as it has only a unique solution and does not involve weights random initialization as training NN does. So an SVM with the similar parameter settings and trained on identical data provides identical results. This increases SVM forecast repeatability while reducing training runs number needed to locate optimum SVM parameter settings [5]. Data non-regularity enables SVMs to be used for regression analysis, for example when data is not distributed regularly or has a known distribution [6].

III METHODOLOGY

Support Vector Machines (SVM) is a learning algorithm that has the unique ability to provide function estimation. Support Vector Regression (SVR), the use of SVMs for regression, operates in high-dimensional feature space to approximate unknown functions in output space, thereby using nonlinear functions to linearly estimate an unknown function. For $y \in \Re, \forall i \in [1, N]$, we have two inequalities that bound the output points of the function to be estimated:

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 10, October 2013

one for the upper boundary and one for the lower boundary. Suppose that we are given a training set:

$$\Omega_r = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), ..., (\mathbf{x}_N, y_N)\}$$

Where $X_i \in \Re^n$ and $y_i \in \Re, \forall i \in [1, N]$.

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. Support vector regression [7] is the natural extension of large margin kernel methods 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 its training 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.

If Support Vector Regression is to be applied to a real-time camera stream, or for estimating several facial features at a time, it is not practical. Lee et al. proposes a ε-smooth SVR [8] formulation, where they only need to solve a system of linear equations iteratively instead of solving a convex quadratic program or a linear program, as is the case with a conventional ε-SVR. Second, they propose a reduction of the kernel, similarly to classification. Those reduced vectors in the kernel are however a subset of the training data. A full SVR is too time-consuming to use for all image locations, and the face space for all poses is too complex to first decide with a classifier where the faces are located. Therefore, we will adapt the approaches to reduce the complexity for classification mentioned above to regression. Both stages, the classification and the regression, are adjustable in their complexity.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 minimizing || 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

$$w^T \emptyset(x_i) - y_i \leq \varepsilon + \xi_i$$
$$y_i - w^T \emptyset(x_i) \leq \varepsilon + \xi_i^*$$
$$\xi_i, \xi_i^* \geq 0, \text{ i= 1,....m}$$

Algorithm for SVR

1. Parameters: η, ε, C.

2. Initialize: $\alpha^* = 0$, $\alpha = 0$.

3. For i = 1… l

$$d_i = y_i - f(\mathbf{x}_i)$$
$$\Delta\alpha_i^* = \eta(d_i - \varepsilon)$$
$$\Delta\alpha_i = -\eta(d_i + \varepsilon)$$
$$\alpha_i^* = \lfloor \alpha_i^* + \Delta\alpha_i^* \rceil$$
$$\alpha_i = \lfloor \alpha_i + \Delta\alpha_i \rceil$$

4. If training has converged stop, else repeat step 3.η is a learning rate parameter.

Regression Analysis

Regression analysis is a statistical tool for the investigation of relationships between variables. Usually, the investigator seeks to ascertain the causal effect of one variable upon another—the effect of a price increase upon demand, for example, or the effect of changes in the money supply upon the inflation rate. To explore such issues, the investigator assembles data on the underlying variables of interest and employs regression to estimate the quantitative effect of the causal variables upon the variable that they influence. The investigator also typically assesses the "statistical significance" of the estimated relationships, that is, the degree of confidence that the true relationship is close to the estimated relationship.Regression is a statistical technique to determine the linear relationship between two or more variables. Regression is primarily used for prediction and causal inference. In its simplest (bivariate) form, regression shows the relationship between one independent variable (X) and a dependent variable (Y). Regression thus shows us how variation in one variable co-occurs with variation in another. What regression cannot show is causation; causation is only demonstrated analytically, through substantive theory. It is important to recognize that regression analysis is fundamentally different from ascertaining the correlations among different variables. Correlation determines the strength of the relationship between variables, while regression attempts to describe that relationship between these variables in more detail.

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 10, October 2013

Regression analysis can be used to identify the line or curve which provides the best fit through a set of data points. This curve can be useful to identify a trend in the data, whether it is linear, parabolic, or of some other form. Regression analysis can be performed using different methods. Regression analysis is used when two or more variables are thought to be systematically connected by a linear relationship. In simple regression, we have only two – let us designate them x and y – and we suppose that they are related by an expression of the form y = b0 + b1 x + e. We'll leave aside for a moment the nature of the variable e and focus on the x - y relationship. y = b0 + b1 x is the equation of a straight line; b0 is the *intercept* (or *constant*) and b1 is the *x coefficient*, which represents the slope of the straight line the equation describes. To be concrete, suppose we are talking about the relation between air temperature and the drying time of paint.

In the Regression Model, the assumptions are, the relation between x and y is given by y = b0 + b1 x + e e is a random variable, which may have both positive and negative values, so

ε is normally distributed

E(e) = 0, the standard deviation of e, syx, is constant over the whole range of variation of x. This property is called "homoscedasticity." Since E(e) = 0, we're supposing that E(y) = b0 + b1x + E(e) = b0 + b1x. Finding the regression line: the method of "ordinary least squares" or OLS, begin with assumed values for b0 and b1 and suppose that the relation between x and y is given by y = b0 + b1x; some b0's and b1's will give us better fits than others. Let $y = a + bx_i$ be the value of y estimated by the regression equation when x has the value xi; then if y_i is actual value, $y_i - \hat{y}_i$ is called the *residual* or the *error*, substituting, let $e_i = y_i - \hat{y}_i = y_i - b_0 - b_1x_i$, different b_0's and b_1's will cause each e_i to have a different value: The residuals along the line marked A are larger than those along the line marked B but the *sum* of deviations is always zero. square each residual and define the sum of squared errors as $\sum (y_i - b_0 - b_1x_i)^2$, x and y are data: the variables are b_0 and b_1, and choosing different values of these will change the size of the sum of squares. Minimizing the sum of squares with respect to b0 and b1, using minimization methods from differential calculus, gives unique values for the b's Resulting formulas are rarely used explicitly anymore, but

$$b_1 = \frac{(\Sigma x_i y_i) - n \times \bar{x} \times \bar{y}}{(\Sigma x^2) - n \times \bar{x}^2}$$

$$b_0 = \bar{y} - b\bar{x}$$

Regression analysis is used to detect a relation between the values of two or more variables, of which at least one is subject to random variation, and to test whether such a relation, either assumed or calculated, is statistically significant. It is a tool for detecting relations between hydrologic parameters in different places, between the parameters of a hydrologic model, between hydraulic parameters and soil parameters, between crop growth and water table depth, and so on.

Regression analysis with MMRE formula (Mean Magnitude Relative Error) and MdMRE (Median Magnitude Relative Error)

It is necessary to measure software estimates accuracy for evaluation and validation. A common evaluation criteria in software engineering [9] is used in this context:

Magnitude Relative Error (*MRE*) computes absolute error percentage between actual and predicted efforts for reference samples.

$$MRE_i = \frac{|actual_i - estimated_i|}{actual_i}$$

The mean magnitude of relative error, MMRE, is the de facto standard evaluation criterion to assess the accuracy of software prediction models. MMRE is a summary statistic, i.e., a single number, aggregating the fundamental metric MRE, a relative residual error. MMRE is used for two kinds of assessments (at least). One purpose of MMRE is to select between competing prediction models: The model that obtains the lowest MMRE is preferred. Another purpose is to provide a quantitative measure of the uncertainty of a prediction (Where a low MMRE is taken to mean low uncertainty or inaccuracy). MMRE calculates *MREs average* over all reference samples. As *MMRE* is sensitive to an individual outlying prediction, a median of *MREs* is adopted for *n* samples (*MdMRE*) when there are many observations less sensitive to extreme *MRE values.* Median Magnitude of Relative Error, defined as *median* over test set of*(|Predicted Effort – Actual Effort| /Actual Effort).* Despite the use of *MMRE* for estimation accuracy, there exists much discussion about its efficacy in estimation procedures. *MMRE* has been criticized as being unbalanced in many validation circumstances, resulting often in overestimation [10].

$$MMRE = \frac{1}{n}\sum_{i=1}^{n} MRE_i$$

$$MdMRE = \underset{i}{median}\left(MRE_i\right)$$

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 10, October 2013

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 wood pulp is largely used for making paper, data of the demand and supply of pulpwood were collected for over ten years for forecasting.

Year	Demand (MT)	Supply (MT)
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

Table 1: sample normalized data of the demand and supply of pulp wood (Metric Tonnes)

The Mean Magnitude Relative Error (MMRE) and Median Magnitude Relative Error (MdMRE) are evaluated through technique like SVM with RBF kernel . Table 1 provides the demand and supply data for TNPL. Table 2 provides results of average MMRE and MdMRE for the SVM with Radial Basis Function (RBF) technique used.

Technique Used	MMRE	MdMRE
SVM-RBF	0.400824	43.92262

Table 2 MMRE and MdMRE for SVM - RBF

V CONCLUSION

This study uses Mean Magnitude Relative Error (MMRE) and Median Magnitude Relative Error (MdMRE) as evaluation criteria for forecasting. The forecasting can further be improved by using optimization technique. 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.

ACKNOWLEDGEMENT

I would like to thank the National Agricultural Innovation Project (NAIP) - A Value Chain on Industrial Agroforestry in Tamil Nadu. Sincere thanks to Mr. R. Sreenivasan, General Manager, Tamil Nadu News print and Papers Limited (TNPL), karur, Dr. K. T. Parthiban, Professor and Head, Tree Breeding, Th. P Durairasu, Dean, FC&RI, Dr. M. Anjugam, Professor and Head, Forest College and Research Institute for their guidance and support.

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] Sansom, D.C., T. Downs, T.K. Saha, "Evaluation of support vector machine based forecasting tool in electricity price forecasting for Australian national electricity market participants. Journal of Electrical and Electronics Engineering, Australia", 22(3):pp.227-234, 2003
[6]Zhang, L., F. Lin, B. Zhang, " Support vector machine learning for image retrieval. In Image Processing, 2001. Proceedings. 2001 International Conference on (2: 721-724). IEEE, 2001
[7] Anandhi, V., &Chezian, R. M, "Support Vector Regression to Forecast the Demand and Supply of Pulpwood", International Journal of Future Computer and Communication, Vol. 2, No. 3, pp.266-269,June 2013
[8] Ratsch, M., et al, "Wavelet Reduced Support Vector Regression for Efficient and RobustHead Pose Estimation ", Ninth Conference on Computer and Robot Vision pp 260-267, 2012
[9] Setiono, R., K. Dejaeger, W. Verbeke, D. Martens, B. Baesens, " Software effort prediction using regression rule extraction from neural networks, In Tools with Artificial Intelligence (ICTAI)", 22nd IEEE International Conference on (2: 45-52). IEEE, 2010
[10] Bhatnagar, R., V. Bhattacharjee, M.K. Ghose, "Software Development Effort Estimation–Neural Network Vs. Regression Modeling Approach", International Journal of Engineering Science and Technology, 2(7): 2950-2956, 2010