10.2478/v10048-010-0011-9

Prediction of Spirometric Forced Expiratory Volume (FEV_1) Data Using Support Vector Regression

A. Kavitha[1], C. M. Sujatha and S. Ramakrishnan[1]

[1]Madras Institute of Technology, Dept. of Inst. Engg., Anna University Chennai, 600044, Chennai, India, ramki@mitindia.edu
College of Engineering, Guindy Campus, Dept. of ECE., Anna University Chennai, 600025, Chennai, India,

In this work, prediction of forced expiratory volume in 1 second (FEV_1) in pulmonary function test is carried out using the spirometer and support vector regression analysis. Pulmonary function data are measured with flow volume spirometer from volunteers (N=175) using a standard data acquisition protocol. The acquired data are then used to predict FEV_1. Support vector machines with polynomial kernel function with four different orders were employed to predict the values of FEV_1. The performance is evaluated by computing the average prediction accuracy for normal and abnormal cases. Results show that support vector machines are capable of predicting FEV_1 in both normal and abnormal cases and the average prediction accuracy for normal subjects was higher than that of abnormal subjects. Accuracy in prediction was found to be high for a regularization constant of C=10. Since FEV_1 is the most significant parameter in the analysis of spirometric data, it appears that this method of assessment is useful in diagnosing the pulmonary abnormalities with incomplete data and data with poor recording.

Key words: Spirometry, forced expiratory maneuver, support vector regression

1. INTRODUCTION

SPIROMETRY is a relatively simple, noninvasive and the most widely used pulmonary function test that measures the volume of air expelled from fully inflated lungs as a function of time. Spirometry measures the volume of air inhaled or exhaled as a function of time during forced breathing maneuvers and is an essential tool in the diagnosis of airway obstruction, and in the detection of restriction and follow-up of respiratory diseases [1]–[4]. It generates a flow-volume signal which represents tidal, inspiratory and expiratory phases of breathing. The significant parameters measured using spirometer are forced vital capacity (FVC), forced expiratory volumes in one second (FEV_1), ratio of FEV_1 to FVC (FEV_1%), peak expiratory flow and forced expiratory flow at 25–75% of FVC (FEF25–75%) . These measured variables are the basis of the diagnosis and treatment of lung disorders [5].

FEV_1 is the volume of air that is forcibly exhaled in the first second, whereas FVC is the total volume of air exhaled after a full inspiration. Airflow obstruction can be diagnosed using spirometry alone by demonstrating a low FEV_1/FVC ratio. A low spirometric FVC together with a normal or high FEV_1/FVC ratio has been classified as a restrictive abnormality [6, 7]. FEV_1 is the most significant parameter for identifying both the restrictive and obstructive respiratory diseases. The value of FEV_1 is very essential in quantifying airflow limitation. It is also a powerful predictor of increased risk of lung cancer and cardiovascular diseases. The American Thoracic Society has developed a scale to rate the severity of disease based on the values of FEV_1 [3].

Spirometry is an effort dependent test that requires the cooperation between the subject and the examiner, and the results obtained will depend on technical as well as personal factors. Also, full inspiration or expiration is difficult to achieve for subjects who suffer with difficulties in breathing. It has been shown in the literature that 50% of the results obtained from spirometric measurement were unacceptable due to failure to complete the test [8]. Since spirometry is the most widely used screening test to investigate the pulmonary function abnormalities, there is a requirement that a large database is to be analyzed by a physician. The spirometric data would also have missing values and patients with lung abnormalities might not be able to repeat the test for acquiring the missing data [9]. Hence, there is a need for prediction of FEV_1 which is the most significant parameter that helps in defining the risk of pulmonary complication in cases where measurements fail to record the parameters.

The prediction of FEV_1 has already been carried out using Radial basis function neural networks [10, 11]. However, conventional neural network methods have demonstrated difficulties finding a good generalization performance. Recently it has been shown that Support vector machines (SVM) have been efficiently employed in various prediction and classification problems. The support vector machines were proposed by Vapnik. The principle of SVM is to find a maximum margin hyperplane for classification by mapping the instances to a higher dimensional space using the kernel function. Kernel function maps the input to a higher dimensional space without computing all elements, which reduces computational complexity and connects the input space and the higher dimensional space directly. SVM then choose a maximum soft margin separating hyperplane in this higher dimensional space, which separates the training instances by their classes [12] – [14].

SVM can be applied to both classification and regression. In the case of classification, an optimal hyperplane is found that separates the data into two classes, whereas in the case of regression a hyperplane is to be constructed that lies close to as many points as possible. Support vector regression is different from conventional regression techniques because it uses structural risk minimization (SRM) but not empirical risk minimization (ERM) induction principle which is equivalent

to minimizing an upper bound on the generalization error and not the training error. Due to this feature it is expected to perform better than conventional techniques which may suffer from possible overfitting. SVM have other desirable properties such as efficient solutions, relatively few adjustable parameters and the interchangeable use of kernel functions [15] –[18]. In this work, an attempt has been made to predict FEV_1 using support vector machines and the results are validated using the average prediction error statistics.

2. METHODOLOGY

For the present study, 175 adult volunteers (normal = 55, abnormal = 50, validation=70) are considered. Spirometric measurements are done with volumetric transducer as it has proven accuracy and stability. The acceptability and reproducibility criterion were adopted as per the recommendation given by the American Thoracic Society [19]. The parameters obtained from the spirometer are subjected to support vector machines. The training dataset included all the measured parameters and the details of the parameters obtained are provided elsewhere [10]. Support vector regression with polynomial kernel is employed in the prediction of FEV_1 values. Regression estimates a function based on a given set. Hence, given a set of data

$$G = (x_i, a_i)_{i=1}^{N}$$

where x_i is the input vector, a_i is the actual value, and N is the total number of data patterns, the SVM regression function is

$$y = f(x) = w_i \varphi_i(x) + b$$

where $\varphi_i(x)$ is the feature of inputs x, and both w_i and b are coefficients which are estimated by minimizing the regularized risk function,

$$R(C) = C\frac{1}{N}\sum_{i=1}^{N} L_{\varepsilon}(d_i, y) + \frac{1}{2}\|w\|^2$$

where,

$$L_{\varepsilon}(d_i, y_i) = \begin{cases} |d_i - y_i| - \varepsilon, |d_i - y_i| \geq \varepsilon \\ 0, else \end{cases}$$

is called the ε - insensitive loss function, d_i is the actual value at period i and y_i is the estimated value at period i and

$$\frac{1}{2}\|w\|^2$$

is the norm of the governing equation. The term C is the regularization constant which specifies the trade-off between the empirical risk and the model flatness. An optimal desired weights vector of the regression hyperplane is represented as

$$w^* = \sum_{i=1}^{N}(\alpha_i - \alpha_i^*) K(x_i, x_j)$$

where α_i and α_i^* are the Lagrangian multipliers. And the regression function is

$$f(x, \alpha_i, \alpha_i^*) = \sum_{i=1}^{N}(\alpha_i - \alpha_i^*) K(x_i, x_j) + b$$

Here, $K(x_i, x_j)$ is called the kernel function. The value of the kernel equals the inner product of two vectors x_i and x_j in the feature space $\varphi(x_i)$ and $\varphi(x_j)$, i.e., $K(x_i, x_j) = \varphi(x_i) . \varphi(x_j)$.

An inner product in feature space has an equivalent kernel in input space,

$$K(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle$$

The polynomial kernel [20, 21] of the form,

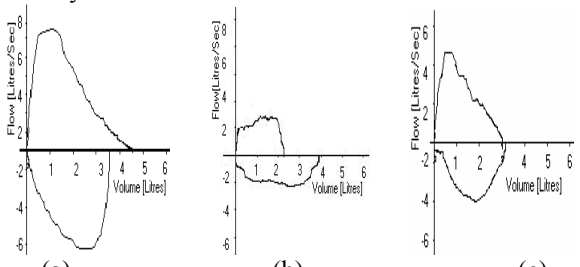
$$K(x, x') = \langle x, x' \rangle^d$$

is employed in the prediction and the performance of the regression model is analyzed by varying the order d of the polynomial.

In this work, the most significant parameter FEV_1 is predicted for 70 test data by training the support vector machine with 105 spirometric data. The regularization constant C is varied from 1 to 10 and the accuracy in prediction of FEV_1 is estimated. Average prediction accuracy for normal and abnormal subjects is calculated for various orders of the kernel function performance of prediction and is evaluated by calculating the accuracy in the prediction of FEV_1. The accuracy in prediction is compared for three chosen values and regularization constants.

3. RESULTS AND DISCUSSION

The typical responses of a spirometer showing the variation of airflow with lung volume for normal, obstructive and restrictive subjects are presented in Fig. 1(a), 1(b) and 1(c) respectively.



(a) (b) (c)

Fig.1 Typical flow volume curves of normal (a), obstructive (b) and restrictive (c) subjects

The normal flow-volume curve is found to have a rapid peak expiratory flow rate with a gradual decline in flow. In restrictive subjects, the shape of the flow volume loop is relatively unaffected, but the overall size of the curve appears smaller when compared to normal. And in obstruction, there was a rapid peak expiratory flow but the curve descends more quickly than normal and takes on a concave shape.

The most significant spirometric respiratory parameter FEV_1 is predicted using support vector regression model employing polynomial kernel. The efficiency of the SVM model is analyzed by varying the regularization constant from 1 to 10. The accuracy in prediction for all the subjects is found to be high for C=10 compared to all the other values. Also, the accuracy is found to saturate after this value and thus further results are presented for C=10. Hence, for the remaining study, C= 10 is chosen to be the optimum upper bound.

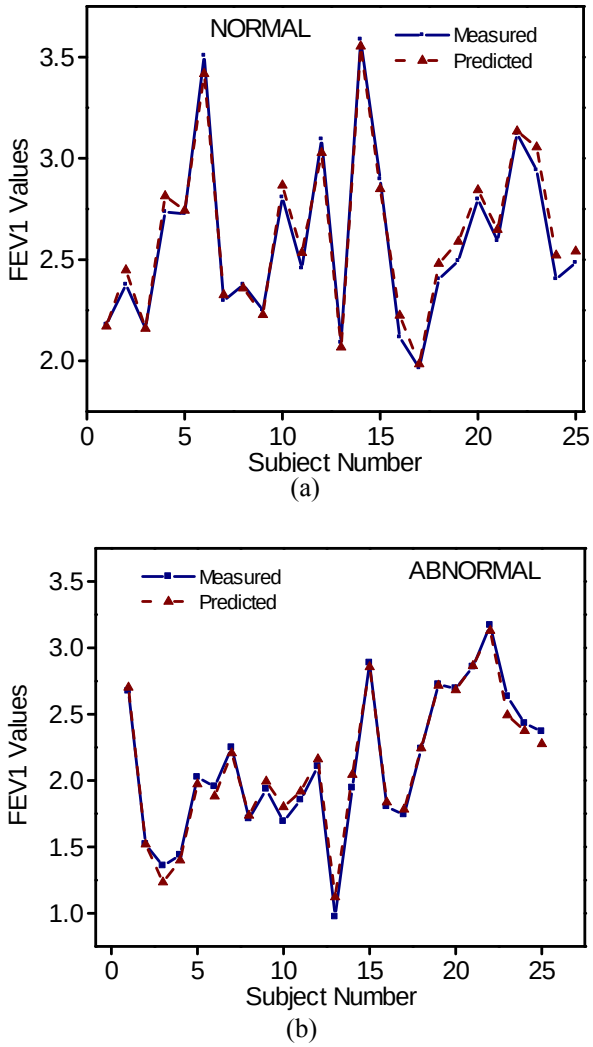


Fig.2 Deviation of the predicted values from the measured FEV_1 values for (a) normal and (b) abnormal subjects

Also, it is found that prediction is efficient irrespective of the subjects and the deviation of predicted values from the measured values is less for normal subjects. Fig. 2(a) and (b) show the deviations in the predicted FEV_1 values from the measured values for normal and abnormal subjects respectively. It is observed from the results that the measured and the predicted FEV_1 values are nearly the same for most of the subjects.

The average prediction accuracies for normal and abnormal subjects are estimated with polynomial kernel with chosen orders. The average prediction accuracy for normal subjects is shown in Fig. 3 as function of order of the kernel for a regularization constant of C=10. It is understood from the results that normal subjects have a consistent accuracy with higher values for order of the kernel to be 2.

Table 1
Statistics of the measured and predicted FEV_1 for all the subjects

Subjects	Mean ± Standard deviation	
	Measured FEV_1	Predicted FEV_1
Normal	2.59±0.43	2.62±0.41
Abnormal	2.12±0.55	2.12±0.53

The statistical analysis of the measured and predicted values for all the subjects is shown in Table 1. It is seen that the mean of both the values for normal subjects is higher than that of abnormal subjects and their standard deviation is also lower.

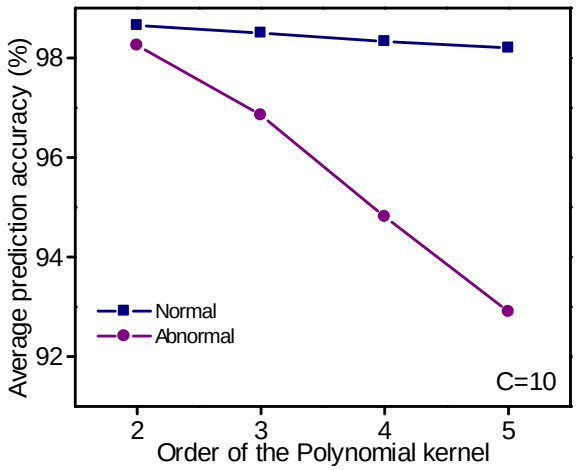


Fig.3 Variation of percentage average prediction accuracy with chosen orders of the kernel for normal and abnormal subjects and regularization constant (C=10)

The number of support vectors employed by the SVM model for the prediction of FEV_1 in normal and abnormal subjects with its corresponding prediction errors is shown in Fig. 4.

It is observed from the results that minimum number of support vectors is utilized by kernel of order 2 when compared to the other orders. Further, the average prediction accuracy is found to correlate well with the number of support vectors employed in the prediction. For a support vector regression

model using a smaller number of support vectors, the prediction accuracies are also found to be high irrespective of the subjects. And the accuracy in prediction is found to be high for normal subjects compared to the abnormal subjects.

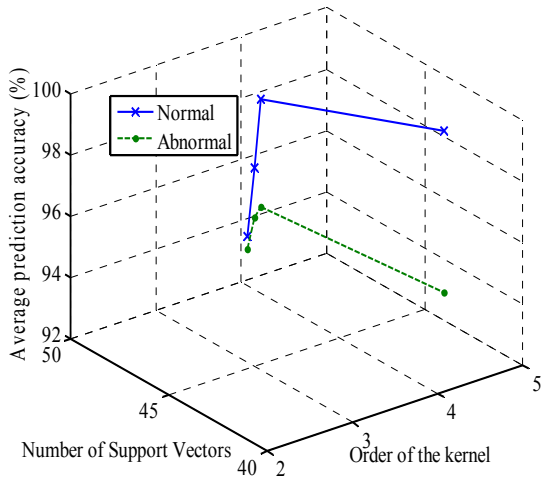


Fig.4 Variations in the average prediction error for normal and abnormal subjects with the number of support vectors employed by the polynomial kernels at chosen orders of the kernel and regularization constant of C=10.

4. Conclusions

Forced expiratory volume in one second is a very useful index for clinical monitoring and assessment. It has been shown that many times spirometric measurements result in incomplete dataset. Hence, the prediction of FEV_1 has been considered significant. There are some earlier reports where prediction of FEV_1 has been demonstrated using neural networks [10].

In this work, attempt has been made to predict FEV_1 values using support vector regression in order to enhance the spirometric investigations. For this analysis both normal and abnormal subjects were used. Support vector regression technique with polynomial kernel was employed in the prediction. Optimum value of the order of the kernel was chosen based on the accuracy in prediction. Further, the support vectors that characterize the regression results were also analyzed. The optimum value of the order of the polynomial kernel was found to be 2 and the regularization constant was 10. The prediction accuracy for normal subjects is found to be high when compared to the abnormal subjects. Further, it seems that optimization of the upper bound of the regularization constant and the kernel width using standard optimization techniques would yield better prediction of FEV_1 with maximum accuracy. We have observed that the degree of severity and classification of abnormalities could be achieved by trying the prediction process with other kernels.

Acknowledgement

The authors would like to thank Dr. Sridharan for his help in clinical data collection.

References

[1] Miller, M.R., Hankinson, J., Brusasco, V., Burgos, F., Casaburi, R., Coates, A., Crapo, R., Enright, P., van der Grinten, C.P.M., Gustafsson, P., Jensen, R., Johnson, D.C., MacIntyre, N., McKay R., Navajas, D., Pedersen, O.F., Pellegrino, R., Viegi, G. and Wanger J.(2005). *Standardisation of spirometry. European Respiratory Journa*l, 26, 319–338.

[2] Pierce, R. (2004). Spirometer: An essential clinical measurement. *Australian Family Phyician*, 34, 535 – 539.

[3] American Thoracic Society. (1991). Lung function testing: selection of reference values and interpretative strategies. *American Reviews on Respiratory Diseases,* 144: 1202-18.

[4] Wagner, N. L., Beckett, W. S. and Steinberg, R. (2006). Using Spirometry results in occupational medicine and research: common errors and good practice in statistical analysis and reporting, *Indian Journal of Occupational Environmental Medicine*, 10, 5-10.

[5] David, P. J., Pierce, R. (2008). Spirometry – The measurement and interpretation of ventilatory function in clinical practice. *Spirometry Handbook*, 3rd edition. 1-24.

[6] Aaron, S. D., Dales, R. E. and Cardinal. P. (1999). How accurate is spirometry at predicting restrictive pulmonary impairment. *Chest*, 115, 869–873.

[7] Sahin, D., Ubeyli, E.D., Ilbay, G., Sahin, M. and Yasar, A. B. (2009). Diagnosis of airway obsruction or restrictive spirometric patterns by multiclass support vector machines, *Journal of Medical Systems*, DOI 10.1007/s10916-009-9312-7.

[8] Ulmer, W.T. (2003). Lung function - Clinical importance, problems and new results. *Journal of Physiology and Pharmacology*, 54, 11-13.

[9] Schermer, T.R., Jacobs, J.E. and Chavennes, N.H. (2003). Validity of spirometric testing in a general practice population of patients with chronic obstructive pulmonary disease (COPD), *Thorax*, 58, 861–866.

[10] Sujatha C. M. and Ramakrishnan S. (2009). Prediction of Forced Expiratory Volume in Normal and restrictive respiratory functions using spirometry and self organizing map, *Journal of Medical Engineering and Technology*, 33, 19-32.

[11] Sujatha C. M., Mahesh V. and Ramakrishnan S. (2008). Comparison of two ANN methods for classification of spirometer data, *Measurement Science Review*, 8 (2), 53 – 57.

[12] Smola A. J. and Schölkopf B. (2004). A tutorial on support vector regression, *Statistical Computing*, 14, 199 – 222.

[13] Vapnik V. N. (1998). Statistical Learning Theory. New York: John Wiley & Sons.

[14] Cristianini, N., and Shawe-Taylor, J. (2000). An introduction to support vector machines. Cambridge: Cambridge University Press.

[15] Lee, J., Blain, S., Casas, M., Kenny, D., and Berall, G.

(2006). A radial basis classifier for the automatic detection of aspiration in children with dysphagia, *Journal of Neuroengineering and Rehabilitation*, 3, 1–17.

[16] Tung-Kuang W., Shian-Chang H. and Ying-Ru M. (2008). Evaluation of ANN and SVM classifiers as predictors to the diagnosis of students with learning disabilities, *Expert Systems Applications*, 34, 1846 – 1856.

[17] Chen K., Kurgan M. and Kurgan M. (2008). Sequence based prediction of relative solvent accessibility using two-stage support vector regression with confidence values, *Journal Biomedical Science Engineering*, 1, 1-9.

[18] Schölkopf B., Kah-Kay S., Christopher J. C. B., Federico G., Partha N., Tomaso P., and Vapnik V. (1997). Comparing Support Vector Machines with Gaussian Kernels to Radial Basis Function Classifiers, *IEEE Transactions on Signal Processing*, 45 (11), 2758 - 2765.

[19] Cooper, B.G. and Madsen, F. (2000). *European Respiratory buyers guide*. 3, 40 – 43.

[20] Hua, S. and Sun Z. (2001). A novel method of protein secondary structure prediction with high segment overlap measure: Support vector machine approach. *Journal of Molecular Biology*, 308, 397–407.

[21] Pai, P. F., Lin, C. H., Hong, W. C. and Chen, C. T. (2006). A Hybrid support vector machine regression for exchange rate prediction, *Information and Management Sciences,* 17, 19-32.

Received December 15, 2009.
Accepted April 6, 2010.