2012 IACSIT Coimbatore Conferences
IPCSIT vol. 28 (2012) © (2012) IACSIT Press, Singapore

Pattern Recognition and Forced Expiratory Volume (FEV1) Spirometric Data Prediction by Support Vector Machine

Prof. Manisha R. Mhetre and Mr Jitendra Khubani

Dept. of Instrumentation & Control Engg, VIT, Pune-411037, Maharashtra, India
E-mail: manisha.mhetre@gmail.com, jitendraanandkhubani@gmail.com

Abstract. A spirometer is an apparatus for measuring the volume of air inspired and expired by the lungs. Spirometry is one of the most widely applied clinical tests in respiratory medicine to diagnose obstructive and to rule out restrictive pulmonary diseases. A key parameter of spirometric lung function tests is forced vital capacity (FVC) and Forced expiratory volume in 1 second (FEV1). In this work, attempt has been made to predict FEV1 values using support vector regression in order to enhance the spirometric investigations. Support vector machine(SVM) constructs a hyperplane or set of hyperplanes in a high-or infinite- dimensional space, which can be used for classification, regression, or other tasks. We have collected data from different hospitals. The acquired data are then used to predict FEV1. Since FEV1 is key parameter in the analysis of spirometric data, so this method is useful in diagnosing the pulmonary abnormalities with incomplete data and data with poor recording. In this paper, we propose an original and universal method by using SVM for prediction of FEV1. We applied the SVM to construct the prediction model and select Gaussian radial basis function (RBF) as the kernel function.

Keywords. Spirometry, FEV1, FVC, Obstructive lung disease, Restrictive lung disease, Support Vector, Regression

1. Introduction :

In machine learning, pattern recognition is the assignment of some sort of output value (or label) to a given input value (or instance). An example of pattern recognition is classification, which attempts to assign each input value to one of a given set of classes (for example, determine whether a given email is "spam" or "non-spam"). However, pattern recognition is a more general problem that encompasses other types of output as well. Other examples are regression, which assigns a real-valued output to each input; sequence labeling, which assigns a class to each member of a sequence of values (for example, part of speech tagging, which assigns a part of speech to each word in an input sentence); and parsing, which assigns a parse tree to an input sentence, describing the syntactic structure of the sentence. Support vector machine (SVM) is a useful technique for data classification, regression and prediction. Previously there has been a lot of study using artificial neural network (ANN) in these areas, especially in the field of prediction.

The foundations of Support Vector Machines (SVM) have been developed by Vapnik and gained popularity due to many promising features such as better empirical performance. The formulation uses the Structural Risk Minimization (SRM) principle, which has been shown to be superior, to traditional Empirical Risk Minimization (ERM) principle, used by conventional neural networks. SRM minimizes an upper bound on the expected risk, where as ERM minimizes the error on the training data. It is this difference which equips SVM with a greater ability to generalize, which is the goal in statistical learning. SVMs were developed to solve the classification problem, but recently they have been extended to solve regression problems. SVM can treat higher dimensional data better even with a relative low amount of training set. Further more, it can present a good ability of generalization for complex model.In this work, attempt has been made to predict FEV1 values using support vector regression in order to enhance the spirometric investigations. For this analysis both normal and abnormal subjects were used.

2. Method Used:

The support vector machine (SVM) is a type of learning machine that is based on statistical theory and it is a popular technique for classification. In order to perform binary deviation, the SVM uses a high dimension space to find a hyperplane where the error rate is minimal. The methodology of SVM can be stated briefly as follows:

We are given a set of training data $\{(\mathbf{x}_1,y_1)...(\mathbf{x}_l,y_l)\}$ in $R^n \times R$ sampled according to unknown probability distribution $P(\mathbf{x},y)$, and a loss function $V(y,f(\mathbf{x}))$ that measures the error, for a given $\mathbf{x}$, $f(\mathbf{x})$ is "predicted" instead of the actual value y. The problem consists in finding a function f that minimizes the expectation of the error on new data that is, finding a function f that minimizes the expected error: $\int V(y,\ f(\mathbf{x}))\ P(\mathbf{x},y)\ d\mathbf{x}\ dy$

For calculating the SVM we see that the goal is to correctly classify all the data. For mathematical calculations we have,

[a] If $Y_i= +1$; $wx_i + b \geq 1$

[b] If $Y_i= -1$; $wx_i + b \leq 1$

[c] For all i; $y_i (w_i + b) \geq 1$

In this equation x is a vector point and w is weight and is also a vector. So to separate the data [a] should always be greater than zero. Among all possible hyper planes, SVM selects the one where the distance of hyper plane is as large as possible.

We need to find w and b such that $\Phi (w) = \frac{1}{2} |w'|\,|w|$ is minimized;

And for all $\{(x_i, y_i)\}$: $y_i (w * x_i + b) \geq 1$.

Now solving: we get that $w = \Sigma\alpha_{i} * x_{i;}$ $b= y_k - w * x_k$ for any x_k such that $\alpha\, k \neq 0$

Now the classifying function will have the following form: $f(x) = \Sigma\alpha_i y_i x_i * x + b$

***SV* classification:**

$$\min_{f,\xi_i} \|f\|_K^2 + C\sum_{i=1}^{l}\xi_i \quad y_i f(\mathbf{x}_i) \geq 1 - \xi_i, \text{ for all i} \quad \xi_i \geq 0$$

Variables ξ_i are called slack variables and they measure the error made at point $(\mathbf{x}_i,y_i)$.

Now we have, $y_i(w'x + b) \geq 1 - S_k$. This allows a point to be a small distance S_k on the wrong side of the hyper plane without violating the constraint. Now we might end up having huge slack variables which allow any line to separate the data, thus in such scenarios we have the Lagrangian variable introduced which penalizes the large slacks.

$\min L = \frac{1}{2} w'w - \sum \lambda_k (y_k (w' x_k + b) + s_k - 1) + \alpha \sum s_k$

Where reducing α allows more data to lie on the wrong side of hyper plane and would be treated as outliers which give smoother decision boundary.

kernels are used to non-linearly map the input data to a high-dimensional space. The new mapping is then linearly separable.

This mapping is defined by the Kernel:

$$K(x,y) = \Phi(x) \bullet \Phi(y)$$

If the feature space is chosen suitably, pattern recognition can be easy.

$$\left\langle x_1 \cdot x_2 \right\rangle \leftarrow K\left(x_1, x_2\right) = \left\langle \Phi\left(x_1\right) \cdot \Phi\left(x_2\right) \right\rangle$$

In this paper, the kernel function of the research is called ***Gaussian Radial Basis Function***

$$K(x, x') = \exp(-\frac{\|x - x'\|^2}{2\sigma^2})$$

For our study, 200 adult volunteers (normal = 70, abnormal = 70, validation=60) are considered. The most significant parameter FEV1 is predicted for 60 test data by training the support vector machine with 140 spirometric data. The prediction of FEV1 has already been carried out using Radial basis function neural networks. However conventional neural network demonstrated difficulties in performance. Support vector machine constructs a hyperplane or set of hyperplanes in a high-or infinite- dimensional space, which can be used for classification, regression, or other tasks. We have collected data from different hospitals. The acquired data are then used to predict FEV1. The performance is evaluated by computing the average prediction accuracy for normal and abnormal cases.

3. Results and Discussions :

FEV_1/FVC (FEV1%) is the ratio of FEV_1 to FVC. In healthy adults this should be approximately 75–80%. In obstructive diseases FEV_1 is diminished because of increased airway resistance to expiratory flow; the FVC may be decreased as well, due to the premature closure of airway in expiration, just not in the same proportion as FEV_1. This generates a reduced value (<80%, often ~45%).

In restrictive diseases the FEV_1 and FVC are both reduced proportionally and the value may be normal or even increased as a result of decreased lung compliance. A derived value of FEV1% is FEV1% predicted, which is defined as FEV1% of the patient divided by the average FEV1% in the population for any person of similar age, sex and body composition.

3.1. Normal flow-curve :

The flow volume curve represents an effort dependent and effort independent part of the curve. All the inspiratory loop and early nearly vertical part of the expiratory loop (peak expiratory flow) are effort dependent that means it depends on muscular effort. This is followed by a linear part of the curve. After about one third of this linear part of the curve is not influenced by any increase in muscular effort and is known as muscle independent part of the loop. This is due to dynamic compression of the airways and the flow of air at this time is due to interaction of the elastic recoil and the airway resistance in the peripheral air passage.

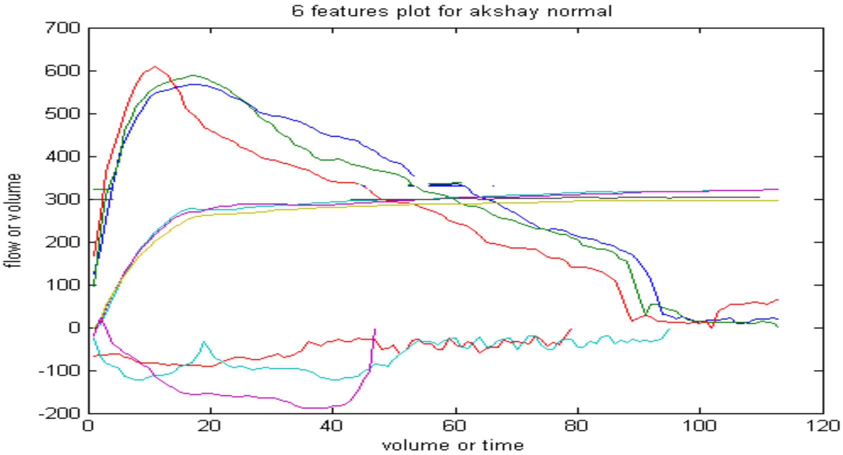

Figure 1: Case Study of Normal Subject (Normal flow v/s volume and volume v/s time loop)

FVC TEST RESULTS FEV1 IS 75% PREDICTED

3.2. Abnormal patterns of flow-volume curve

The shape of the flow-volume curve gives a clear idea whether it records are normal or abnormal. The abnormalities are either due to obstructive or restrictive ventilatory impairment (Figure 2,3).

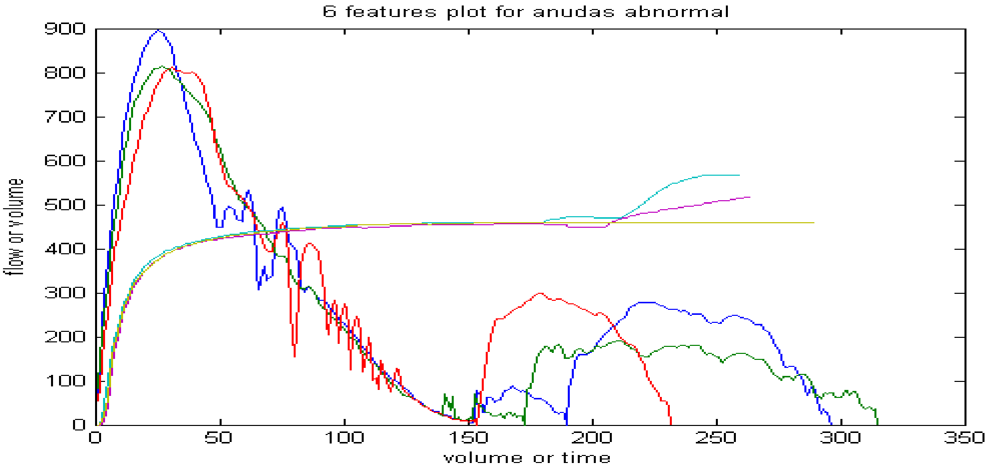

Figure 2: Case Study of obstructive Subject (Normal flow v/s volume and volume v/s time loop)

FVC TEST RESULTS FEV1 IS 87% PREDICTED

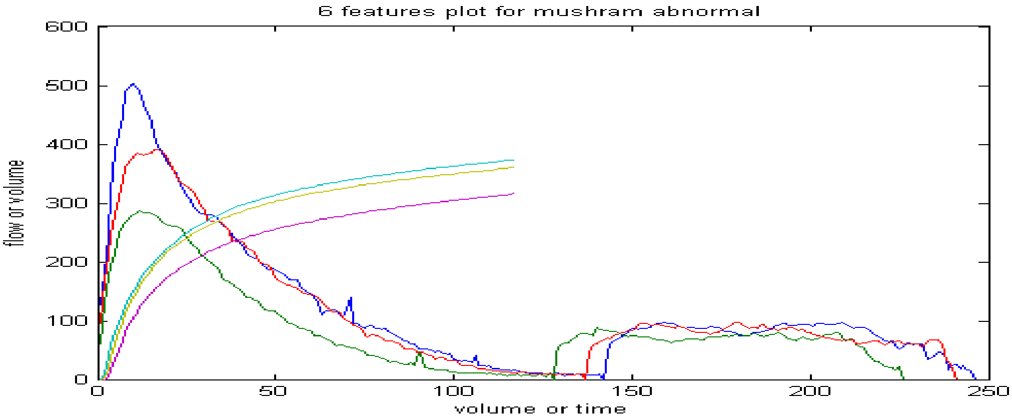

Figure 3: Case Study of Restrictive Subject (Normal flow v/s volume and volume v/s time loop)

FVC TEST RESULTS FEV1 IS 72% PREDICTED

4. Conclusions:

Support vector machines are capable of predicting FEV1 in both normal and abnormal cases. Since FEV1 is key parameter in the analysis of spirometric data, so this method is useful in diagnosing the pulmonary abnormalities with incomplete data and data with poor recording.

5. 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 Journal, 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.