2.4. Support Vector Machine (SVM)

SVM is a supervised learning technique which depends on the training dataset to obtain input-output mapping functions formula. Vapnik (1995) proposed support vector machines approach to solve the problems of regression and obtained good performance for several time series and regression forecasting problems. The concept of mapping function includes either using the classification function which categorize the input datasets, or using the regression model which forecasts the dependent variable.

The origin of SVM is belong to "kernel methods". It also belong to mathematical basics in statistical learning theory. SVM is used in many different real world applications, such as face recognition, data mining, image processing and others along with soft computing techniques such as artificial neural networks ANN and fuzzy logic techniques (Wang, 2005)

SVM can also be applied to nonlinear regression and pattern recognition problems. it is suggested to provide higher performance and more accurate results than classical statistical and learning machines approaches. it can be introduced as useful tool for solving the problem of classification.

SVM is a useful approach for classifying input–output datasets that is easier to use than (ANN). It can be used to solve classification pattern recognition, regression estimation problems, and time series forecasting. It has an advantages due to ability to improve the problem of nonlinear time series and regression that caused by good results of time series forecasting.

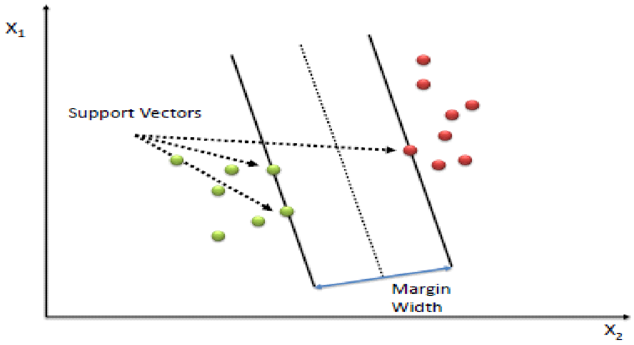
There are several types of learning algorithms, SVM is one of these types which can be use the kernel function and identify the correct decision function. SVM can show good capacity to perform time series forecasting problem with high performance (Ismail & Shabri, 2014).

There are two separated classes in training datasets, and there are a margin that separates between these classes in the dependent variable. The objective of SVM is to formulate the hyperplane (HP) to be as far as possible from the nearest points of two vectors or classes. In other word, SVM achieves perfect classification by formulate HP which maximizes the margin of separating two support vectors.

Support that HP is:

$$wx + b = 0 \qquad (2.29)$$

where w is the margin width and b is the bias of HP from the origin point while $G = \{(x_i, y_i) = 1, 2, \ldots, n\}$ and $x_i \in R^n$ that is the ith input vector $y_i \in \{-1, 1\}$ instead of {0,1} as in computer fields is a categorical target variable (Ganapathiraju *et al.*, 2000; Huang *et al.*, 2005; Singh & Kaur, 2012) as explained in figure2.5 below.



(saedsayad.com)

Figure 2.5 The hyperplane that maximizes the margin between the two classes.

Defining an optimal HP which maximizes the margin of separation represents the first step in the simplest algorithm of SVM. An optimal HP can be defined as the function that maximizes the margin of width (w). Figure2.6 explains SVM in more details.

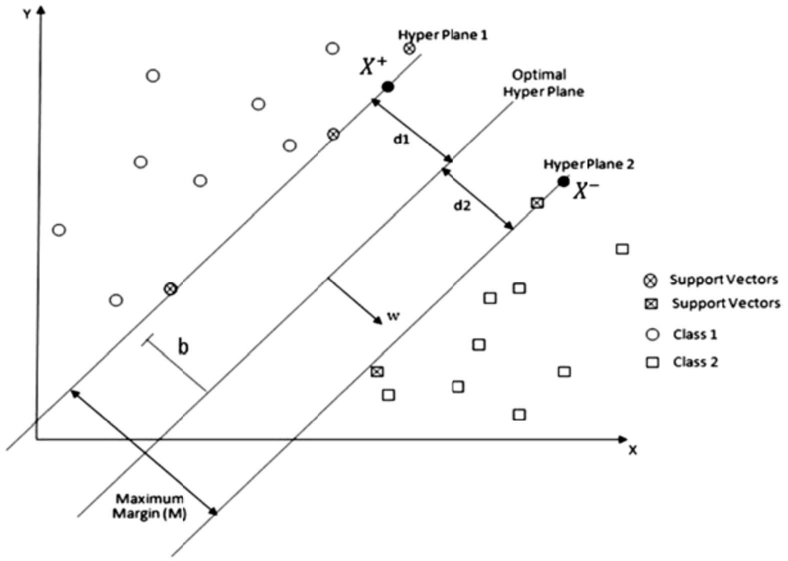


Figure 2.6 Linear separable binary classification (Singh & Kaur, 2012).

Where SVM is used to classify data values by performing the following conditions.

$$(w.x+b)\geq 1 \quad \text{if } y_i = 1, \tag{2.30}$$

$$(w.x+b)\leq -1 \quad \text{if } y_i = -1, \tag{2.31}$$

The points in figure2.6 that lie closest to the separating margins of HP can be called the support vectors; then the two HPs; HP1 and HP2 lie on this points can be formulated such as follows.

For HP1, $w.x_2+b=1$ (2.32)

For HP2, $w.x_1+b=-1$ (2.33)

Therefore, the distance from a point x^+ on HP1 to optimal HP was given by d_1 such as

$$d_1=\frac{1}{\|w\|} \tag{2.34}$$

And the distance from a point x^- on HP2 to optimal HP was given by d_2 such as:

$$d_2=\frac{1}{\|w\|} \tag{2.35}$$

The pooled margin width can be computed as the following :

$$M=\frac{2}{\|w\|} \tag{2.36}$$

For minimizing the quadratic programming (QP) to achieve the optimization, the minimization term will be as follows :

$$\min\frac{1}{2}\|w\|^2 \tag{2.37}$$

s.t.

$y_i \left(w .x_i +b\right) \geq 1$ for all x_i

For solving this type of optimization problem Lagrange's multiplier $(LM)_\alpha$ will be used the saddle point, where $\alpha_i > 0$ for all i; the primal formula of this function is as follows:

$$L_p = \frac{1}{2}\|w\|^2 - \alpha_i \left[y_i \left(w .\mathrm{x}_i + b\right) - 1\right] \qquad (2.38)$$

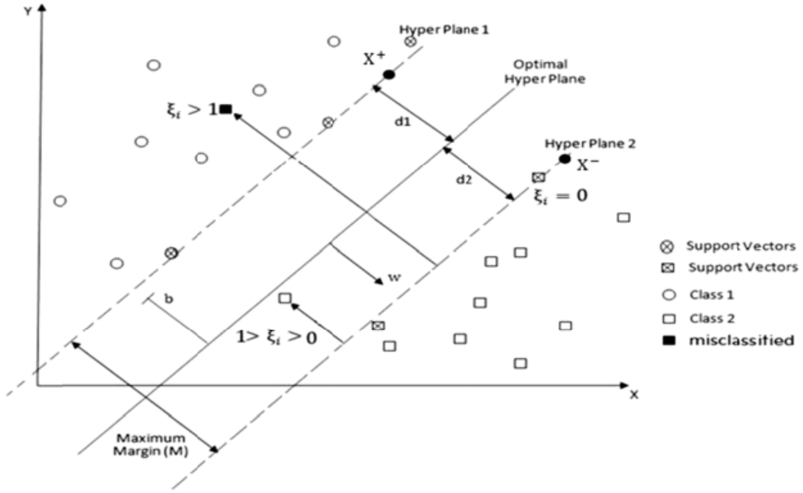


Figure 2.7 Linearly non-separable binary classification (Singh & Kaur, 2012).

The optimal HP can be found by minimizing the objective function such as follows.

$$\min \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N} \zeta_i \qquad i = 1, 2, \ldots, N \qquad (2.39)$$

s.t

$$y_i \left(w .\mathrm{x}_i + b\right) \geq 1 - \zeta_i \qquad \zeta_i > 0 \qquad i = 1, 2, \ldots, N$$

Where ζ_i represents the slack variable and C represents the parameter of regularization which can set a control between using the slack variable ζ_i penalty and changing the size of the margin that means there are several optional choices as follows.

1- When C is small, the constraints of wide margin can be easily ignored which means a lot of sample values that may not lie in an ideal position.
2- When C is large, the constraints of narrow margin can be easily ignored which means few sample values that may not lie in an ideal position.

To find the optimal HP, the objective and classifier function as the following :

$$f(x) = sign\left[\sum_{i=1}^{N} \alpha_i y_i k(x_i, x) + b\right] \qquad (2.40)$$

where α is Lagrange's multiplier $k(x_i, x) = x_i . x$, and N is the number of support vectors, and the sign in this function can determine a sample belong to which class.

2.4.1. Advantages and Disadvantages of SVM

1-They are characterized by being based on simple mathematical ideas.

2-They are of high performance in various practical applications, as they have the ability to process large feature spaces.

3-They deal with complex non-linear problem using a simple linear algorithm using the concept of Kernel Function.

4-In the train of data training, over-training or over fitting can be controlled using the Soft Margin Approach (Abe, 2005).

On the other hand, SVM has disadvantages representing another problem of support vector machines which is in fact slow training. Because support vector machines are trained by solving a quadratic programming problem with the number of variables equal to the number of training data, training is slow for a large number of training data (Burges, 1998).