Using Support Vector Machine To Determine the Limits of Multivariate Control Charts.

H E Fahad ¹, Gh J Asmaa ²
¹Baghdad University, College of Administration & Economics
²Baghdad University, College of Administration & Economics
Email:fahad89_1@yahoo.com

Abstract: Quality control is one of the effective statistical tools in the field of productivity control to monitor and match manufactured products to standard specifications and approved standards for some goods and services, and their purpose is to keep pace with the production and industrial development in the labour market and competition. Naturalness in the productive process. The SVM method was used (Support Vector Machine) It is one of the algorithms that are used to solve classification problems or determine the limits of control in monitoring the production process in the fact that the data do not follow the normal distribution or have an unknown distribution, as the research aims to monitor the production process through several variables at the same time to reflect the quality The material produced. The drinking water test data for the old Shatra water project was used to clarify the method of operation of the backing wave machine method, as the standard drawing method was adopted to indicate the performance and efficiency of the method used, and its performance was good to know the quality of the.

1-Introduction

The issue of statistical control is one of the most important topics that have a prominent role in achieving objective specifications regarding the product and increasing the ability to control the quality of industrial products that have been produced and to monitor defects through the production process. Among the most important methods used in quality control are (Statistical Process Control) (SPC)), which is considered as a means of quality control using statistical methods and monitoring the quality characteristics of manufactured materials to ensure that they adhere to certain standards and maintain the production process in a stable condition, and then follow the progress of the production process and detect non-normal changes that enable us to intervene early to correct the production process.

Control charts are divided into two types: The first type is called univariate control chart, as focusing through it on the product by using only one variable to determine the process that is under control or outside it, assuming that the variables of the production process are independent and this is an incorrect procedure in most cases because most production processes It consists of several variables and with technological advancement and the complexity of industrial processes, the need for the second type has emerged, which requires monitoring many variables, and with the use of computers and information technology, it has become possible to monitor several variables simultaneously to
monitor operations to determine the process if it is under control or outside it and this is called (Multivariate Non Parametric Control Charts) (Multivariate Statistical Process Control) (MSPC) which depend on parameterized multivariate control charts (Parametric Multivariate Control Charts) on the assumption that the basic distribution of the production process is the normal distribution. In fact of the matter, this assumption is not applied in all cases and thus it will give inaccurate results if the data does not follow the normal distribution in the case of applying the parameter methods of the production process [1]. Therefore, the production process cannot be well monitored. To get rid of these problems, one can use; (Multivariate Non Parametric Control Charts), because it is considered a better alternative, hence the research problem includes that most control charts are based on monitoring one characteristic or one variable that reflects the quality of the product. It requires knowledge of the distribution and therefore a method has been applied (Support Vector Machine), as a method for determining the limits of control in the case of multiple variables for surveys that do not follow the normal distribution, by using the standard drawing to clarify the performance of the method used. The importance of non-parametric multivariate control charts is that they do not need to know the basic distribution of the process and are less affected by outliers. Therefore, the use of non-parametric multivariate control charts is useful in monitoring production processes regardless of knowledge of the basic distribution of the process, thus providing an easy option to apply traditional methods of statistical operations [2].

2. Method and Method

2.1. Support Vector Machine (SVM)

This method is one of the Supervised machine learning algorithms used for classification. It was proposed on the basis of Statistical Learning theory [3], it is the idea presented for the first time by the researcher (Vladimir N. Vapnik) in 1963 and in 1995 (Vapnik, V., ed.) [4] was developed, and the SVM technology was originally created to solve pattern recognition issues by specifying the level of separating or optimal solution (hyper plane) for the data to be separated [5]. The hyper plane is defined as a phrase. From the surface, the elements or data are separated into two classes according to the type of data of the distance between the (Hyper plane) and the closest element of any class called the margin. Machine For Classification) as well as it was used to solve other statistical problems such as estimating the parameters of the regression model and prediction, as well as dealing with heterogeneous data, and it was called the Support Vector Machine For Regression. Or not clear correctly, and this is achieved when there is specific data [6] and the Support Vector Machine (SVM) makes the hyper plane as far as possible from it. The two classes are the basic idea of the (SVM) method (Support Vector Machine) that classifies data into dimensions, that is, more than one linear separator, because it is primarily designed to separate two classes from each other, but it can be generalized to multivariate problems. It gives him preference in many diagnostic problems such as bankruptcy or breast cancer diagnosis and other diverse problems and its ability to deal with high and large data sets as well as its ability to address prediction problems. SVM takes the entered data and predicts it based on the teaching data and classifies it into two classes. The larger the surface separating the data, the greater accuracy. It is also distinguished by its ability to detect outliers and can be used in quality control to monitor production processes and to find the best separating surface between data. There are two types [7].

First type: Linear Classifier:
It is a straight line test so that it can separate the data and be as close as possible to this data to determine the different measures of similarity. If the data is two-dimensional (D = 2), the data is divided into two parts [8]. Where (D is the number of dimensions) as in the following figure:
Figures (1) two-dimensional linear classification [9]
But if it is multi-dimensional (D > 2), the data is divided into three straight lines to form the optimal solution, as in the following figure:

Figures (2) represents the linear classification of more than two dimensions.
Type Two: (Non Linear Classifier)
The non-linear classifier in some classification problems does not have a clear separating level so that it can be used as a separator criterion. This is found by using (Kernel) to create a separating surface that can separate the data and be as close as possible to all these data, as shown in the following figure:

Figures (3) represents the distribution of lines for nonlinear classification [9]
The concept of (SVM) support vector machine shares important characteristics with quality control [9], which are as follows:
•(SVM) is a binary classifier that separates data into two classes (positive class and negative class).
The control charts divide data into two classes (inside control and outside control)
Control charts determined the control limit that separates the two states (inside control and outside control). The control limit is also determined by (SVM).

The control charts and (SVM) in general are based on the decision function by using the statistical method, reducing risks and knowing the limits of the statistical control charts in product monitoring.

2.1.1. Formula for the Support Vector Machine (SVM)

Many SVM algorithms are applied in binary classification and converting the binary solution to the optimal first solution using mathematical relationships and this method works well with small and medium-sized data problems. Hyper Plane Maximum Margin separates a set of points, since we assume that we have n points so that each value of $x_i$ is entered has several dimensions (d-Dimensionality) and the values of $y_i$ are equal to one positive or negative one, so the trained data is according to the following.

$y_i = +1$ about the group $y_i = -1$, meaning that $y_i$ is the classification of the classifier $[+1,-1]$ so that the points closest to the optimal solution or the level of separation (Hyper plane) to the origin point ((Support Vector) have a value of either $+1$ for the first group or $-1$ for the second group, as in the following Figure 4 [10], [11]:

![Figure 4](image)

Figure (4) shows the process of separating the dividing line

The separating surfaces can be illustrated by the two equations, as in the above figure

\[
\begin{align*}
w.x_i + b &= 1 \\
w.x_i + b &= -1
\end{align*}
\]

(1) (2)

So that

\[
w.x_i + b = 1 \quad : \text{The vector of weights is represented and the normalization is to solve the optimization or the interval plane}
\]

\[
b: \text{Intercept Tem or limit Bias}
\]

\[
x: \text{is the matrix of explanatory variables with dimensions n * d}
\]

the SVM technique achieves better expectations by maximizing and therefore if we assume that the data can be separated linearly, the (hyper plane) between the two surfaces is $\frac{2}{||w||}$, and since they should not be points inside the margin, the following two conditions have been added to classify the observations[12]

\[
\begin{align*}
w.x_i + b &\geq 1 \quad \text{for } y_i = +1 \\
w.x_i + b &\leq 1 \quad \text{for } y_i = -1
\end{align*}
\]

(3) (4)
So that \( i = 1, 2, \ldots, L \).

By combining the two equations, we get the following condition:
\[
y_i(w \cdot x_i + b) \geq 1 \quad \text{for } i = 1, 2, \ldots, L
\]
\[
y_i(w \cdot x_i + b) - 1 \geq 0 \quad \text{for } i = 1, 2, \ldots, L
\]
(5)

The distance between the support vector machine (SVM) and the separating plane can be defined by the margin of the supporting vector machine (SVM margin) and to make the separating plane (hyper plane) away from the backing vector machine, we need to maximize the margin, and the distance between the secondary levels in equation (2) and (1) is equal to \( \frac{1}{2} ||w||^2 \), since the purpose of the separating plane is to find the values of the two variables \( w, b \) In order to reduce the weights according to the following initial formula
\[
\min \frac{1}{2} ||w||^2
\]
\[
\text{s.t. } y_i(w \cdot x_i + b) \geq 1 \quad \text{for } i = 1, 2, \ldots, L
\]
(7)

In order to verification restrictions of this function, the Lagrange Multipliers \( \alpha \) can be used, so the equation for the Hyper plane can be written as \([10], [13]\)
\[
L(W, b, \alpha) = \frac{1}{2} ||w||^2 - \sum_{i=1}^{L} \alpha_i[y_i(w \cdot x_i + b) - 1]
\]
(9)
\[
L(w, b, \alpha) = \frac{1}{2} ||w||^2 - \sum_{i=1}^{L} \alpha_i w_x_i + b \sum_{i=1}^{L} \alpha_i y_i + \sum_{i=1}^{L} \alpha_i
\]
(10)

We find the value of \( w, b \) by differentiating the formula \( L(w, b, \alpha) \) and making the derivative equal to zero. Then we get the values of the variables \([7]\)
\[
\frac{dL(w, \alpha)}{dw} = w - \sum_{i=1}^{L} \alpha_i y_i x_i
\]
(11)
\[
\frac{dL(w, \alpha)}{dw} = 0
\]
(12)
\[
w = \sum_{i=1}^{L} \alpha_i y_i x_i
\]
(13)
So that
\[
\alpha_i (\text{ Lagrange Multipliers})
\]
\[
0 \leq \alpha_i \leq C
\]
C: is a criterion that works by balancing the estimation error and the amount of divergence from the principal axis
We find the derivative with respect to \( b \)
\[
\frac{dL(w, \alpha)}{db} = \sum_{i=1}^{L} \alpha_i y_i
\]
(15)
We set the derivative with respect to \( b \) equal to zero
\[
\frac{dL(w, b)}{db} = 0
\]
(16)
\[
\sum_{i=1}^{L} \alpha_i y_i = 0
\]
(17)
Any point that equation (17) will be satisfying support vector machine , and we denote it by the symbol \( x \) and it will be written as follows
\[
y(w \cdot x + b) = 1
\]
(18)
Substituting the value of \( w \) into formula (14), we get
\[
y \sum_{i=1}^{L} (\alpha_i y_i x_i + x + b) = 1
\]
(19)
We multiply both sides by \( y \) and the equation after simplification becomes the following
\[ y^2 \sum_{i=1}^{\infty} (\alpha_i y_i x_i x + b) = y \]  

\[ (w, x_i - b) = y \]  

\[ b = y - \sum_{i=1}^{\infty} (\alpha_i y_i x_i x_i x_i x_i) \]  

\[ \frac{1}{N} \sum_{i=1}^{N} y - (\alpha_i y_i x_i x_i) x_i \]  

Thus, we can classify any new point for any of the classification groups according to the following formula

\[ y_i = (w, x_i + b) \]  

Figure (5) shows the processes of observations inside and outside control [9]

3. Experimental results

The experiment application includes the use of the support vector machine method and then the use of the drawing method to demonstrate the efficiency of the method used in detecting the variables inside or outside the boundaries of control. This research was based on data related to drinking water treatments in the old Shatra water project for the year (2017-2019).

Chemical and physical tests have been taken for the most important variables that must be available for the water to be potable, as follows:

1. \( X_1 \): represents (Potassium (K)), is one of the important chemical elements for the human body, animals and plants.
2. \( X_2 \): represents (Sodium (Na)) It is one of the most common chemical elements, which constitutes approximately 80% of the dissolved components in sea water and can be obtained from salts.
3. \( X_3 \): Representing (Total Dissolved Solids(T.D.S)), where these salts come as a result of the melting of some layers of rocks and soil as a result of other effects resulting from household, industrial and agricultural activities.
4. \( X_4 \): represents (Sulphates (\( \text{SO}_4 \))) is a chemical compound consisting of the united atoms of sulfur, oxygen and a lot of sulfates used to solubility in water.
5. \( X_5 \): (Chloride(CL)) is one of the important ions found in nature. It is used to filter drinking water and it is easy to obtain through salts.
6. \( X_6 \): (Magnesium (Mg) is an essential element for the proper functioning of cells in the body.
7. \( X_7 \): (Calcium (Ca)) is an important element in the life of Living organisms.
8. \( X_8 \): (Total Hardness(T.H.E)) describing the condition of the water when the salt content is high.
9. X₉: Turbidity is the presence of particles of mud, impurities, or microorganisms in the water, which leads to turbidity of the water color.

10. X₁₀: Fluoride represents a salt found in nature in places where there is water.

The table below shows the standard specifications for drinking water with the maximum permissible limits and the unit of measurement approved by the Dhiqar Water Directorate.

| The term index | Code | measuring unit | Allowed upper limits |
|----------------|------|----------------|----------------------|
| 1 Turbidity    | Turb | NTU            | 5                    |
| 2 Total Hardness | T.H | Mg/L           | 500                  |
| 3 Calcium      | Ca   | Mg/L           | 150                  |
| 4 Magnesium    | Mg   | Mg/L           | 100                  |
| 5 Chloride     | CL   | Mg/L           | 350                  |
| 6 Sulphates    | S₀₄  | Mg/L           | 400                  |
| 7 Total Dissolved Solids | T.D.S | Mg/L | 1000 |
| 8 Sodium       | Na   | Mg/L           | 200                  |
| 9 Potassium    | K    | Mg/L           | 5                    |

3.1. Test Data

We use the (Kolmogorov – Smirnov (KS test)) as it is one of the nonparametric tests that compares the distribution of a statistical population through random samples taken from this population and can be used to compare the distributions, especially (Normal Distribution)) to find out that the data follow the normal distribution or not and at a significant level (0.05) The test hypothesis can be written as follows:

H₀: Drinking water data distributed a normal distribution
H₁: Drinking water data is not distributed a normal distribution

And by using the program (SPSS: 23) to test the distributions to see if the data are distributing a normal distribution or not, the results of Table (2) show the values of the test power index. (P-Value) for all tests is less than the level of significance (0.05). Therefore, the drinking water test data are not distributed in the normal distribution, and thus the null hypothesis was rejected.

| Variables | Kolmogorov Smirnov | P.Value |
|-----------|--------------------|---------|
| X1        | 0.128              | 0.000   |
| X2        | 0.067              | 0.020   |
| X3        | 0.088              | 0.025   |
| X4        | 0.087              | 0.025   |
| X5        | 0.102              | 0.004   |
| X6        | 0.096              | 0.009   |
| X7        | 0.102              | 0.004   |
| X8        | 0.094              | 0.011   |
| X9        | 0.086              | 0.030   |
| X10       | 0.096              | 0.008   |
Through the figures related to testing the normal distribution of drinking water data, it was found that these data do not match the normal distribution.

**Figure (6)** represents the normal distribution test for the (K) data

**Figure (7)** represents the normal distribution test for the (Na) data.

**Figure (8)** represents the normal distribution test for the (T.D.S) data.

**Figure (9)** represents the normal distribution test for the (SO4) data.
Figure (10) represents the normal distribution test for the (CL) data

Figure (11) represents the normal distribution test for the (Mg) data

Figure (12) represents the normal distribution test for the (Ca) data

Figure (13) represents the normal distribution test for the (T.H) data
3.2. Applied Support Vector Machine Method

This method was applied after writing the program in Matlab: 2018b. The experiment was conducted on samples of drinking water treatments for the old Shatra water project, as the number was 120 observations, and it was of two types: the first type represents the samples that have been processed and the second type represents raw samples (not valid for drinking) as the SVM algorithm was used to divide the samples into two phases of the training phase using a mixed number of the two types (treated and untreated water) for the purpose of training. As for the second phase, it was used for the purpose of testing the accuracy of the classification to find out observations inside or outside the control limit using this method and was calculated. The criteria for the accuracy of the classifier according to certain criteria were calculated based on the following indicators:

1. (True Positive) (TP) Diagnosis of unprocessed sample
2. (True Negative) (TN) Correctly diagnose the treated sample.
3. (False Positive) (FP) Incorrectly diagnosed untreated sample.
4. (False Negative) (FN) Diagnosis of the sample was treated incorrectly.

Based on these indicators, the following set of criteria is calculated:

a. Accuracy: represents the ratio between the results that were correctly diagnosed (Positive, Negative) to the overall results

\[
\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \times 100\% \tag{25}
\]

b. Specificity: represents the ratio between the samples that were diagnosed properly processed (True Negative) to the total samples that were diagnosed properly processed with the samples that were incorrectly diagnosed.

\[
\text{SPE} = \frac{TP}{TN + FP} \times 100\% \tag{26}
\]
c-Sensitivity: represents the ratio between properly diagnosed untreated samples to all untreated diagnosed samples (correctly and incorrectly)

$$\text{SEN} = \frac{TP}{TP+FN} \times 100\%$$  \hspace{1cm} (27)

d-False Positive Rate: is the ratio between untreated and incorrectly diagnosed samples to all samples that were correctly diagnosed (correctly (TN) or incorrectly (FP))

$$\text{FPR} = \frac{FP}{TN+FP} \times 100\%$$  \hspace{1cm} (28)

Where the indicators were applied and the results were as shown in Table No. (3)

| Table (3) Indicators of the support vector machine |
|-----------------------------------------------|
| TN | TP | FN | FP |
| 60 | 60 | 0  | 0  |

The performance specifications of the classifier (SVM) were calculated according to the standards mentioned above and as shown in Table (4)

| Table (4) criterion of the support vector machine |
|-----------------------------------------------|
| Acc | SPE | SEN | FPR |
| 100% | 100% | 100% | 0   |

And based on these criterion and specifications, it is found that the samples are under control and that the (SVM) method isolated the treated and untreated samples well, and this shows that the drinking water treatments for the Shatera water project are within the specifications according to the (SVM) method and as shown in Figure (16), which shows the lack of Any observation from the treated samples came out of control, while the untreated observations went out of control. This indicates that the SVM method is highly efficient.

![Figure (16) shows the control chart for the SVM method](image)

**Conclusions**

Through the results, the SVM method showed that no observations came out of the tested data. This indicates the accuracy of the method by testing the specifications of potable water and it gives accurate and efficient results and shows that the observations are within the limits of control.
References

[1] Lowry, C.A., Woodall, W.H., Champ, C.W. and Rigdon, S.E.,( 1992). " A multivariate exponentially weighted moving average control chart", Technometrics, Vol. 34, No.(1), pp.46-53.

[2] Zou, C. and Tsung, F., (2011). "A multivariate sign EWMA control chart", Technometrics, Vol. 53, No.(1), pp.84-97.

[3] Ge S., Gao Y. and Wang R., (2007). "ACM,Least Significant Bit Steganography Detection with Machine Learning Techniques ", National Laboratory for Novel Software Technology Nanjing University 210093 Nanjing, Jiangsu, China.

[4] Vapnik, V., ed., (1995)."The nature of statistical learning theory", New York, Springer, pp.1-201

[5] Ivanciue, O., (2007)."Applications of support Vector Machines in Cjemistry", Wiley-VCH, John Wiley & Sons, Volume 23, pp. 291-400.

[6] Duan J. C., Härdle, W.K., Gentle J.(2012)."Handbook of computational finance", Series: Springer Handbooks of Computational Statistics, Publisher: Springer-Verlag Berlin Heidelberg, ISBN: 978-3-642-17253-3, 978-3-642-17254-0.

[7] Camci, F., Chinnam, R. B., and Ellis, R. D.,(2008)."A robust kernel-distance multivariate control chart using support vector principles", International Journal of Production Research, Volume 46, Issue 18, pp 5075-5095

[8] Ahmad. A., Khalid. M. and Yusof R.(2002.("Machine Learning Using Support Vector Machines", Universiti Teknologi Malaysia, pp.1-8, abdraham@uniten.edu.my.

[9] Gani, W., Taleb, H., and Limam, M.,(2008)."Statistical Process Control using Support Vector Machines: A Case Study", Taken from The International Journal for Quality and Standards, ISG, University of Tunis, pp.1-17

[10] Fletcher, T. (2009)."Support vector machines explained", University College London, London, pp.1-18

[11] Moradi, M., Sardasht, M. S., and Ebrahimpour, M.,(2012)."An Application of Support Vector Machines in Bankruptcy Prediction; Evidence from Iran", World Applied Sciences Journal, Vol. 17, No. 6, pp. 710-717, ISSN 1818-4952

[12] Tax, D Avid. M.J., Duin, Robert. P.W.,(2004)."Support Vector Data Description", Machine Learning, Vol. 54, PP.45–66

[13] Lee, M.C. & To, C.,(2010)."Comparison of Support Vector Machine and Back Propagation Neural Network in Evaluating the Enterprise Financial Distress", International Journal of Artificial Intelligence & Applications (IJAIA), Vol.1, No.3, pp 31-43