International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 1 Jan 2022, pp: 117-123 www.ijaem.net ISSN: 2395-5252

Machine Learning Algorithms in Fault Diagnosis of a Process Plant

*[1]Madu, C., [1,2]Folami, N.A. and [1]Onoriode, E.

[1]Department of Chemical Engineering, Lagos state Polytechnic, Ikorodu Lagos Nigeria [2]Department of Petroleum and Gas Engineering, School of Science Engineering and Environment, University of Salford, Manchester, United Kingdom

Submitted: 01-01-2022 Revised: 05-01-2022 Accepted: 10-01-2022

ABSTRACT
Machine learning algorithms and data monitoring and reduction machines are useful in handling large data generated during measurements in process plants. These machines are used for extraction of useful features from these large accumulated data and for fault classification. This paper has presented a number of these machine algorithms that are trained for data handling, treatment and fault classification. They are now mostly hybridized for process control and fault diagnosis
Key Words: Data, Classification, Training, Information Transfer, PDF, Fuzzy Clustering

I. INTRODUCTION

There are several methods of fault detection and diagnosis. They are classified into model-based methods, data driven methods and knowledge-based methods. Because of thedifficulty in obtaining exact process mathematical model of a system,the other methods are now being used more often. The data driven methods apply statistical techniques for fault detection and diagnosis (He et al, 2017). Many process control laboratories have large data archives during normal and faulty operating conditions. These large data can be used to build a statistical model which can be used to detect the faults in the future. The larger the data the more accurate will be the results given by the statistical model (Isermann, 2004). Most of the machine learning algorithms are data classifiers.Some machine learning algorithms such as PCA and PLS are used for data reduction and monitoring in addition to fault diagnosis.

Principal Component Analysis(PCA)

All types of scientific data analysis have the collection of observations on a physical or social phenomenon. The complexities of most phenomena require observations on more than one variable to be collected and, therefore, most data can be characterized as exhibiting multivariable behaviour. When more than one variable is observed, some form of correlation will exist between individual variables. Multivariable analysis simultaneously investigates all the variables to reveal the relationships between them, in order to interpret the data very well (Papazoglou 1998).

Relationships can range from independence to collinearity. PCA is a multivariate statistical analysis method. It is defined as a linear transformation of the original variables, which are normally correlated into a new set of variables, which are uncorrelated or orthogonal to each other (Wang et al 2004).Reducing the dimensionality of a problem by removing some of the variables is done using PCA. When reduction is performed by PCA, the original variables can be represented by a smaller number of principal components because of the redundancy of the variables.

The concept of latent variable is the technique used in the reduction of dimensionality of multivariable data. A latent variable is a hypothetical variable constructed for the purpose of understanding a characteristic of interest that cannot be measured directly (Papazoglou 1998). They are not observable; however, they have a certain impact on the measured variable, that is why they are subject to analysis. PCA uses the concept in its data reduction technique. Harold Hotalling in 1933, proposed PCA for analyzing the covariance and correlation structures between a number of random variables. PCA puts multivariable data sample containing significant redundancies, in terms of a set of uncorrelated latent variables, each of which is a linear combination of the original variables.

In PCA, we start with a multivariable sample of observations, which characterizes n objects with respect to the random variables $x_1, x_2 \ldots x_m$ and which is represented by a data matrix X of dimension (n x m).The latent variables

International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 1 Jan 2022, pp: 117-123 www.ijaem.net ISSN: 2395-5252

are also called principal components. The number of principal components is less than or equal to the original variables. The first principal component has the largest possible variance which means, it accounts for as much of the variability in data as possible. Principal components are orthogonal to each other. The resulting vectors are an uncorrelated orthogonal basis set (Wikipedia). The principal components are orthogonal because they are the eigen-vectors of the covariance matrix, which is symmetric.

Partial Least Squares Model (PLS)

Partial least squares regression model can be used in the same way similar to PCA for process monitoring. PLS works by simultaneously decomposing both the input data block X and the output data block Y (Wise et al ND). Y is a vector if there is only one output variable. The decomposition is done in such a way that the factor scores in X and Y blocks have the maximum covariance. The number of factors, also known as latent variables, retained in the PLS regression model is optimized based on prediction through a series of cross-validations (Wise et al ND). The parameters used in PLS prediction can be reduced to a single linear equation.

$\hat{Y} = XB$ (1)

where B is a matrix. In the case of single output variable, B is a vector. Then

$\hat{Y} = X\sum_{i=1}^{k} b_i\left[\prod_{j=1}^{i-1}\left(I - w_j P_j^T\right)\right]w_i q_i^T$ (2)

b_i are the inner relation coefficients, w_j and the P_j are the X block weights and loadings and the q_i are the Y block loadings?When PLS is used to monitor processes in a way similar to PCA, then nPLS models would be required. However, with Equation (2), the n PLS models can be formed into a single matrix with each model being a column vector. PLS is an optimization of these regression models to improve their predictability.

Classification Methods

A classifier is a machine learning model that is used to discriminate different objects on certain features.A classifier is an algorithm itself – the rules used by machines to classify data. A classification model is the end result of classification of data by machine learning. The model is trained using the classifier, so that the model, ultimately classifies the data. There are both supervised and unsupervised classifiers. When a machine learning classifier is fed with unlabeled datasets which it classifiers according to pattern recognition or structures and faults in the data, it is called unsupervised machine learning. Supervised and semi-supervised classifiers are fed with training datasets, from which they learn to classify data according to predetermined categories.

c-Means clustering methods

Clustering algorithms are used to assign unlabeled data (data for which the class of operation is unknown) to one of c classes, where c is two or more. When training data are labelled, it means the class of operation of each data point is known a priori. Data points clustering using a hard c- means clustering algorithm have crisp membership functions (House et al, 1999, Bezdek 1981). That is, if data are clustered into two classes, each data point will have a membership of unity in one class and zero in the other. In addition, each of the c classes can be represented by a single prototype data point, which is typically just the mean value of all members of the class.

Also, a fuzzy c means clustering algorithm can be used. Data clustered using this algorithm have fuzzy membership functions. Here, each data point will have membership ranging from zero to unity in each class, with the sum of the membership values for a given data point being unity and the sum of the membership values for a given class being greater than zero and less than the total number of clustered points n. Data points that fall midway between the cluster centers tend to have nearly equal membership in each class. The fuzzy c means algorithm produces prototype data points for each of the c classes; but the prototype data points are weighted mean values of the class members. The membership functions obtained from the hard c-means and fuzzy c means clustering algorithms can be used in nearest neighbor algorithms to classify new data points. The c prototype data points can be used in nearest prototype algorithms. The membership functions and prototype can also be used to train an artificial neural network algorithm. Fuzzy membership functions and prototype data points are determined by minimizing the objective function J_m (Bezdek, 1981).

$J_m(U,V) = \sum_{j=1}^{n}\sum_{i=1}^{c}\left(u_{ij}\right)^m \left\|x_j - v_i\right\|^2$ (3a)

where x_j is the jth q-dimensional training data point (for data characterised by seven residual values, q =7), v_i is the ith q-dimensional fuzzy prototype data point, u_{ij} is the membership of the jth training data point in the ith class, U is a matrix of membership functions containing a c-dimensional membership function for each of the n clustered points, v is a matrix containing a q dimensional fuzzy prototype data point for each of the c classes, m is any real value greater than or equal to unity, and. ‖.‖ is the 2-norm, defined by

International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 1 Jan 2022, pp: 117-123 www.ijaem.net ISSN: 2395-5252

$$\|x_j - v_i\| = \sqrt{\sum_{l=1}^{q}(x_{jl} - v_{il})^2} \quad (3b)$$

From Equation (3a) it is seen that the larger the distance between x_j and v_i, the smaller the membership function u_{ij}. If m = 1. J_m simplifies to the hard c-means objective. As m increases, J_m becomes more insensitive to distance and membership functions become more- fuzzy.

k-Nearest Neighbour classifier

This is used for pattern recognition (Bezdek 1981, Schalkoff 1992, House, et al, 1999). It is based on the fact that data from a specified class of operation should fall within the same region of feature(residual) space. Using training data that has been assigned either a crisp or a fuzzy membership label (or membership function), the distance from a test data point to each training data point is determined. The training data points are then sorted based on their distance to the test point. Classification of the test point is performed by computing the average membership function of the k- nearest neighbours (k training data points closest to the test point) in the training data set and assigning the test point to the class having the largest average membership function value. The main disadvantage of this method is in training data storage and computational efficiency, especially when large training data are involved.

k-Nearest Prototype Classifier

It is similar to k-NN type classifier discussed above. However, in k-nearest prototype classifier k-NP, the training data used in the k-NN classifier are replaced by a prototype data points representing each of classes of operation. For 1-NP classification, a test point is assigned to the class of operation of the closest prototype data point. This can be extended to a more general case of multiple prototypes for each class of operation. For1-NP classification, prototype data points have membership in a single class of operation (crisp membership). For k-NP classification, prototype data points can have either crisp or fuzzy membership functions. We can see here that in 1-NP classification method computation difficulties is alleviated as seen in k-NN, however some information in the training data may be lost. If there is an overlap of classes, a data point would likely be improperly classified. The k-NP is a compromise between the two methods.

(i) Artificial Neural Network Classifier

This expert system is very good in mapping inputs to outputs. It is especially very useful for nonlinear systems. For the purpose of classification, we employ a feedforward neural network and train it to produce a specific output pattern for specific input pattern. We can use patterns of residuals (Saif et at, 2021, Berrar et al, 2018) as inputs to the NN classifier and crisp membership functions representing various classes of operation as outputs. Alternatively, training data can be input directly to an NN and the output be crisp or fuzzy membership functions. Fuzzy prototypes and membership functions would again come from a fuzzy c mean algorithm. However, NN are black boxes where the reasoning behind decisions may be difficult to comprehend. They also don't extrapolate, which means, input patterns unlike those used for training may produce unreasonable outputs (Srivastava et al,2014, Ahmed et al, 2016)

Artificial neural network uses large historical data for its training and classification. It is one of the best machine learning algorithms available. It can be put to use in the following areas:

- Signal processing: suppress line noise, with adaptive echo cancelling, blind source separation
- Dynamic and static process modelling, trend analysis which requires a lot of data.
- Nonlinear and adaptive control; Manufacturing plants for controlling automated machines.
- Quality prediction and control
- Fault detection and diagnosis
- Multivariable pattern recognition
- Data validation and rectification.
- Time series prediction
- Robotics - navigation, vision recognition
- Pattern recognition, i.e. recognizing handwritten characters, e.g. Apple's Newton uses a neural net
- Medicine, i.e. storing medical records based on case information
- Speech production: reading text aloud (NETtalk)
- Speech recognition
- Process optimization
- Automated decision making
- Vision: face recognition, edge detection, visual search engines
- Business, e.g. rules for mortgage decisions are extracted

Bayes Classifier

This uses probability function. A Bayes classifier minimizes the cost or the probability of misclassification. That means, it is an optimum classifier, classifying an observation x_o to one of c classes (House, 1999). The cost of misclassification is minimized if x_o is allocated to the class k that

International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 1 Jan 2022, pp: 117-123 www.ijaem.net ISSN: 2395-5252

minimizes the following expression (John and Wichern 1992).

$\sum_{\substack{i=1 \\ i\neq k}}^{c} P_l f_i(x_o) C(k \mid i)$ (4)

Where P_i is a priori probability of an observation coming from class i, f_i,is the conditional density function for class i. and $C(k \mid i)$ is the cost associated with allocating an observation to class k, when in fact it comes from class i. If the cost of misclassification $C(k \mid i)$ is the same for all i and k, the decision rule states that $x_o$should be allocated to the class k that minimizes

$\sum_{\substack{i=1 \\ i\neq k}}^{c} P_i f_i(x_o)$ (5)

This expression is minimum when $P_k f_k(x_o)$ is maximum. The Bayes decision rule therefore states.

allocatex_o to class k if

$\text{In}P_k f_k(x_o) > P_i f_i(x_o) \ \ for\ i \neq k$ (6)

Equivalent expression is

allocatex_o to class k if

$\text{In}P_k f_k(x_o) > InP_i f_i(x_o) \ \ for\ i \neq k$ (7)

If the data in each class can be represented by a multivariate normal density function,

$f_i(x) = \frac{1}{(2\pi)^{q/2}|\sum_i|^{1/2}} exp\left[-\frac{1}{2}(x-\mu_i)' \sum_i^{-1}(x-\mu i\right.$ (8)

whereμ_i is the mean vector and $\sum_i$ is the covariance matrix, Eq. (7) becomes:

allocatex_o to class k if

$d_k^Q(x_o) = largest\ of\ d_1^Q(x_o), d_2^Q(x_o), \ldots, d_c^Q(x_o)$ (9)

whered_k^Q denotes the quadratic discrimination score (or quadratic score) for class k given by

$d_k^Q(x_o) = -\frac{1}{2} ln|\sum_k| - \frac{1}{2}(x_o-\mu_k)' \sum_k^{-1}(x_o-\mu k + lnPk - q2 ln(2\pi)$ (10)

The final term in Eq. (10) is the same for all classes and can be dropped. The quadratic score for class i can be approximated by replacing the population mean μ_i and covariance matrix $\sum_i$ with the sample mean $\bar{x}_i$ and the covariance matrix S_i

When all the covariance matrices are taken to be equal, Eq. (10) simplifies to

$d_k^Q(x_o) = -\frac{1}{2}(x_o-\mu_k)' \sum_k^{-1}(x_o-\mu_k) + lnP_k$ (11)

The first and last terms in Eq. (10) have been dropped because they are the same for all classes. The quadratic score of class i can be approximated by replacing the population statistics μ_i and $\sum$ with the sample mean$\bar{x}_i$ and a pooled estimate Spooled of $\sum$ given by

Spooled $= \frac{(n_1-1)S_1+(n_2-1)S_2+\cdots+(n_c-1)S_c}{n_1+n_2+\cdots n_c-c}$ (12)

wheren_i is the sample size for class i.

We see here that Bayes classifier is an optimal classifier. However, its performance is sensitive to departures from normality in the data.

The simple form of Bayes classifier is the naïve Bayes classifier. Naïve Bayes is a family of probabilistic algorithms that calculates the possibility that any given data point may fall into one or more group of categories or not. In a non-scientific application for example, naïve Bayes is used to categorize customers comments, news, articles, email etc. into topics or tags to organize them according to their predetermined criteria. The crux of the Bayes classifier is based on the Bayes theorem. (Butcher et al, 2008)

Using Bayes theorem, one can find the probability of A happening, given that B has occurred. Here, B is the evidence and A is the hypothesis. It uses the assumption that the variables / features are independent. That means that the presence of one particular feature does not affect the other.

Types of naïve Bayes algorithm include: Bernoulli naïve Bayes which uses some Boolean type variables – true or 'false', or 'yes' or 'no'. It is used when the multivariate data is distributed according to Bernoulli distribution. Multinormal naive Bayes algorithm is used to solve document classification problems. It is a sorting type algorithm which uses the frequency of the present words as features. For example, documents could be sorted into legal category or non-legal category. The Gaussian naïve Bayes is used for continuous value data which is assumed to distribute according to Gaussian distribution.

Relevance vector machines

The relevance vector machine (RVM) is a Bayesian-based machine learning method proposed by Tipping (2001). It is adopted as a predictor in reliability prediction. Such predictions include bearing prediction, battery reliability prediction, software prediction and process equipment prediction. RVM can also estimate the posterior probability of predicted object at each prediction step. When applied for prediction of equipment, the value of the characteristic index as well as its probability density function (PDF) can be obtained at the same time (Zhu et al, 2021). Based on the predicted PDF and the preset failure threshold, the operational reliability of the equipment can be obtained. Principal component analysis is then applied to combine several characteristic indexes into one hybrid index to increase the robustness and accuracy of the operational reliability prediction.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 1 Jan 2022, pp: 117-123 www.ijaem.net ISSN: 2395-5252

The steps of the method include the following. First of all, some representative characteristic indexes of running equipment are obtained, and PCA is applied to these indexes to get the hybrid index. Next, the series of the hybrid index of long term monitoring is used to train a single step prediction RVM model to predict the value and probability density function (PDF) of the next step. Finally, the operational reliability of the equipment is calculated by the interval integral defined by the failure threshold and the predicted value of the hybrid index.

In this method, performance degradation information is the only thing required, hence, the method is suitable for reliability estimation problems.

Fuzzy Logic Engine.

This is another machine learning algorithm that uses IF and THEN rules to classify data. It is used in process control and fault diagnosis. It is applied in single and hybrid form.

A fuzzy logic engine has four principal components.

1. Fuzzification: that is, conversion of non-fuzzy (Crips) input data into suitable logistic values which is the labels of the fuzzy sets.
2. Fuzzy rule base, which consist of a set of linguistic diagnostic rules written in the form

IF a set of conditions are satisfied, THEN a set of consequences are inferred

3. Fuzzy inference machine, which is a decision-making logic that employs rules the fuzzy rule base to infer fuzzy diagnostic actions in response to fuzzy inputs.
4. Defuzzification, which converts the range of values of output variables into corresponding universe of discourse

Steps in using the fuzzy logic machine.

1. Determination of the objective and criteria for the diagnosis
2. Determination of the input and output relationship and chosen a minimum number of variables
3. Construction of a rule base
4. Creation of membership function that defines the meaning (values) of input/output terms used – the rules.

Creation of pre-and-post processing fuzzy logic routine or programming the rules into fuzzy logic engine

Fuzzy clustering.

Clustering is the allocation of data points to a certain number of classes (Tracy 2013). Each class has a cluster center. This is the point which best presents the data in the cluster. The aim of fuzzy clustering is that each data point belongs to all classes with a certain degree of membership. Now the degree to which a data point belongs to a certain class is dependent upon the distance to all cluster centers.

Each class could correspond to a particular fault.

The techniques are as follows:

- Offline phase: Learning phase, consists of the determination of the characteristics (i.e. cluster centers) of the classes. Here, use a learning data set containing residuals for all known faults.
- Online phase: Calculates the membership degree of the current residuals to each of the known classes.

A fuzzy clustering algorithm called the c-means, minimizes the following convergent criterion.

$J = \sum_{i=k}^{n} \sum_{i=1}^{c} (\mu_{i,k})^m \, d_{i,k}$ subject to: $\sum_{i=1}^{c} \mu_{i,k} = 1$ (13)

c donates the number of clusters, n = number of data, $\mu_{i,k}$, the fuzzy membership of the k-th point to the i-th cluster, $d_{i,k}$, the Euclidean distance between the data point and the cluster center, m, a fuzzy weighting factor which defines the degree of fuzziness of the results.

In general, m=2 is chosen (Tracy, 2013). The distance $d_{i,k}$ is obtained as follows:

$(d_{i,k})^2 = |k_k - v_i|^2$ (14)

v_i is the cluster center, v is defined as the fuzzy weighted center of gravity of the data.

$$v_i = \frac{\sum_{i=1}^{n} (\mu_{i,k})^m}{\sum_{k=1}^{n} (\mu_{k,i})^m} \quad (15a)$$

$\mu_{k,i}$ is the partition matrix (matrix of membership functions) which denote the extent to which the data point x_k is similar to each cluster center. The solution which minimizes J_m are:

$$\mu_{j,k} = \frac{1}{\sum_{j=1}^{c} \left(\frac{d_{k,i}}{d_{k,j}} \right)^{2/m-1}} \quad (15b)$$

Procedure to calculate the fuzzy c-means. The algorithm is as follows

1. Choose the number of classes $2 \leq c < n2$; choose m, $1 \leq m < \infty$. Initialise U^0
2. Calculate the cluster centers, v_i using Eq.3
3. Calculate new partition matrix $U^{(1)}$ using Eq.4

International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 1 Jan 2022, pp: 117-123 www.ijaem.net ISSN: 2395-5252

4. Compare $U^{(1)}$and $U^{(1+1)}$. If the variation of membership degree $\mu_{i,k}$ calculated with an appropriate norm, is smaller than a given threshold, stop the algorithm, otherwise go back to step 2.
The determination of the cluster centers is then complete.

Relationship between fuzzy reasoning and fuzzy clustering

If someone has full expert knowledge about the system then fuzzy reasoning can be used. If not available, but what is there is fault free and fault containing data, then fuzzy clusters can be used since it is data-based method.

If both expert knowledge and data exists, then expert knowledge is used to form an initial rule base, and then data is used to determine the membership function parameters.
Suppose there are two inputs A and B, and suppose 3 cluster centers have been generated, C1, C2, and C3.
A set of corresponding rules for the system is, e.g. if cluster C1, corresponds to fault free, cluster C2 to fault 1 and cluster C3 to fault 2, then the rule becomes
IF U1 is A2 AND U2 is B2 THEN fault free
IF U1 is A1 AND U2 is B3 THEN fault 1
IF U1 is A1 AND U2 is B2 THEN fault

Multiscale entropy and support vector machine

Entropy is a measure of uncertainty of a process. This was successfully applied in thermodynamics (Pan et al, 2016). Multiscale entropy (MSE) is now used in fault diagnosis (Lin et al, 2010, cited by Pan et al, 2016). The algorithm uses variation of parameters. To avoid redundancy when this algorithm is used and thus reduce training time, mutual information technique was introduced to it (Doquire et al, 2013, cited by Pan et al, 2016) to select the most effective features from the extracted entropies to improve efficiency. Once a fault vector has been selected, a multifault classifier is required to identify the fault conditions. The support vector machine (SVM) is one of such machine learning algorithms for this purpose. SVM has a high degree of accuracy and good generalization.

To use the above algorithm, first calculate the sample entropy values from the given data of the process. This is done using sample entropy and multiscale entropy. In the second step, mutual information technique is used to select the effective entropy features. The last step is to use the support vector machine, which is a pattern classifier, to discriminate the faults.

Sample entropy and multiscale entropy are basically used for feature extraction from large scale data from measurements especially from time series data.
Sample entropy can be summarized as follows (Pan et al 2016)
Let X be a time series of length N
$X = [x(i), x(i+1), \dots, x(i+m-1)]$
(16)
Step 1 Construction of template vectors
$X_i^m = [x(i), x(i+1), \dots, x(i+m-1)]$, i = 1, 2, N-m (17)
Step 2 When the distance between two template vectors (X_i^m, X_j^m) is smaller than predefined tolerance r, a match has occurred. The distance between the two vectors is given by
$d(X_i^m, X_j^m) = max_{k=0,1,\dots,m-1}[|x(i+k) - x(i+j)|$ i,j,= 1,2,…,N-m, $i \neq j$ (18)
Step 3 Suppose $n_i^m(r)$ is the number of distances within r and $B_i^m(r)$is the total number of m=dimension matched vector pairs then:
$B_i^m(r) = \frac{n_i^m(r)}{(N-m-1)}$ I = 1, 2,…,N-m (19)
Step 4 Define the average value of $B_i^m(r)$ such that

$$B^m(r) = \sum_{i=1}^{N-m} \frac{B_i^m(r)}{N-m)}$$

Step 5 Repeat steps (1)-(4) for m+1, to obtain $B^{m+1}(r)$
For finite N, Sample En is defined as the logarithm of the ratio of $B_i^m(r)$ and $B^m(r)$
$SampEn(m, r, N) = -In\frac{B^{m+1}(r)}{B^m(r)} = InB^m(r) - InB^{m+1}(r)$ (20)
N is the length of the time series, m is the dimension of the sequences to be compared, and r is tolerance. Tolerance is best set at r x SD (Pan et al 2016), where Sd is the standard deviation of the data set
Multiscale entropy: The regularity in time series is referred to as granularity. The MSE is made up of:
A coarse graining step: This is used to obtain the representation of the original time series on different scales, and the SampEn step which is used to quantify the regularities of the coarse grain time series. A coarse grain series in MSE is given by
$y_j = \frac{1}{\tau}\sum_{i=(j-1)r+1}^{n} x(i)\ 1 \leq j \leq \frac{N}{\tau}$ (21)
When $\tau = 1$, the coarse grain time series is the original time series, but as τ increases, the length of the resulting coarse-grained time series decreases.

Mutual information: The extracted SampEn and MSE features are now able to discriminate different types of faults. It is now required to reduce the features containing the

International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 1 Jan 2022, pp: 117-123 www.ijaem.net ISSN: 2395-5252

faultMI is a measure of the variables independence. This will improve the performance of the vector support machine classification and hence the fault diagnosis. Mutual information (MI) I(X, Y) is the amount of uncertainty in X due to the knowledge of Y. This obtained as:

$$I(X,Y) = \sum_{x,y} p(x,y)\log\frac{p(x,y)}{p(x)p*y)} \quad (22)$$

where p(x,y) is the joint probability distribution function of X and Y, p(x) and p(y0 are their marginal probabilities.

Support vector machine (SVM) This machine is capable of handling large feature space and the dimension of the classified does not have influence on the performance. It has a better generalization property, it can also be used in nonlinear classification with the selection of proper kernel functions (Pan et al 2016. For the training data, the two-classclassification problem is solved as follows:

$$\text{minimize: } \frac{1}{2}\left\|w^{ij}\right\|^2 + c\sum_t \xi_t^{ij}\,(\omega^{ij})^T$$

$$\text{Subject to } (\omega^{ij})^T\varphi(x_t) + b^{ij} \geq 1 - \xi_t^{ij} \quad (23)$$

$$(\omega^{ij})^T\varphi(x_t) + b^{ij} \geq \xi_t^{ij} - 1$$

$$\xi_t^{ij} \geq 0$$

c is a penalty parameter of error classification.

II. CONCLUSION

All the algorithms mentioned above when properly trained can be used for fault diagnosis in large process plants where several measurements generate lot of data. The data generated are usually used to extract features and the classification of the faults can be carried out based on the extracted features.

REFERENCES

[1]. Ahmed, D.F. an Khalaf, A.H. (2016). Artificial NNs controller for crude oil distillation col.Of Baiji Refinery. J.Chem.& Process Techn. Vol.7, issue 1.
[2]. Bezdek, J.C. (1981). Pattern recognition with fuzzy objective function algorithms.
[3]. New York: Plenum Press
[4]. Berrar, D. (2018) Bayes' theorem and naïve Bayes' classifier. Encyclopedia of Bioinformatics and computational Biology, vol1, Elsevier, pp403-412
[5]. Butcher, S.G.W., and Sheppard, J.W. (2008).Distributed Smoothing in Bayesian fault diagnosis. IEEE Transactions on instrumentation and measurement. Vol xx, No x.
[6]. Doquire et al, (2013), cited by Pan et al 2016.
[7]. He, Y., Kusiaki, A., Ouyang, T. and Teng, W. (2017). Data-driven modelling of truck engine exhaust valve failures: A case study. J of Mech. Science and Techn, vol31, No 6, pp2747-2757.
[8]. Isermann, R. (2004) Model based fault detection and diagnosis, status and applications
[9]. IFAC, pp 11-22.
[10]. Papazoglou, M. (1998).Multivariable statistical process control of chemical processes.PhD thesis, University of Newcastle upon Tyne.
[11]. Pan, S., Han, T., Tan, A.C.C. and Lin, T.R. (2016). Fault diagnosis system of induction motors based on mutiscal entropy and support vector machine with mutual information algorithm
[12]. Hindwi Publishing corporation, vol 2016, I.D 5836717
[13]. House, J.M., Lee, W.Y., Shin, D.R.(1999). Classification techniques for fault detection and diagnosis of an air handling unit.ASHRAE Transactions.Symposia.
[14]. Tracy, D. (2013). Fuzzy logic in fault diagnosis.Dept of Measurement and control, Gehard Molecular Universitat.G.H. Duisberg, Germany.
[15]. Saif, S., Das, p. and Biswas, S. (2021). A hybride model based on MBA-ANFIS for COVID19 confirmed cases prediction and forecast. J Inst Eng. India.
[16]. Schalkoff, R.J. (1992). Pattern recognition: Statistical structure and and neural approaches. NY: John Wiley and sons
[17]. Srivastava, N.P., Srivastava, R.K. and Vashishtha, P.K. (2014). Fault detection and isolation (FDI) via NN. J of Engrg Research and Application vol. 4 issue 1, (version 1), pp81-86
[18]. Wise, B.M. and Ricker, N.L. (ND). Recent advances in multivariate statistical process control: Improving robustiness and sensitivity.
[19]. Zhu, L., Chen, D and Feng, P. (2021) Equipment operational reliability evaluation method based on RVM and PCA fused features. Hindwi Mathematical problems in Engineering, vol. 2021, ID 6687248