Contents lists available at GrowingScience

International Journal of Industrial Engineering Computations

homepage: www.GrowingScience.com/ijiec

# Feature-based decision rules for control charts pattern recognition: A comparison between CART and QUEST algorithm

**Monark Bag**[a], **Susanta Kumar Gauri**[b] and **Shankar Chakraborty**[a*]

[a] *Department of Production Engineering, Jadavpur University, Kolkata – 700 032, India*
[b] *SQC & OR unit, Indian Statistical Institute, 203, B. T. Road, Kolkata – 7000108, India*

ARTICLEINFO

*Article history*:
Received 30 April 2011
Received in revised form
September, 01, 2011
Accepted 01 September 2011
Available online
2 September 2011

Keywords:
*Control chart pattern*
*Shape feature*
*Pattern recognition*
*Decision tree*
*CART*
*QUEST*

ABSTRACT

Control chart pattern (CCP) recognition can act as a problem identification tool in any manufacturing organization. Feature-based rules in the form of decision trees have become quite popular in recent years for CCP recognition. This is because the practitioners can clearly understand how a particular pattern has been identified by the use of relevant shape features. Moreover, since the extracted features represent the main characteristics of the original data in a condensed form, it can also facilitate efficient pattern recognition. The reported feature-based decision trees can recognize eight types of CCPs using extracted values of seven shape features. In this paper, a different set of seven most useful features is presented that can recognize nine main CCPs, including mixture pattern. Based on these features, decision trees are developed using CART (classification and regression tree) and QUEST (quick unbiased efficient statistical tree) algorithms. The relative performance of the CART and QUEST-based decision trees are extensively studied using simulated pattern data. The results show that the CART-based decision trees result in better recognition performance but lesser consistency, whereas, the QUEST-based decision trees give better consistency but lesser recognition performance.

© 2012 Growing Science Ltd. All rights reserved

## 1. Introduction

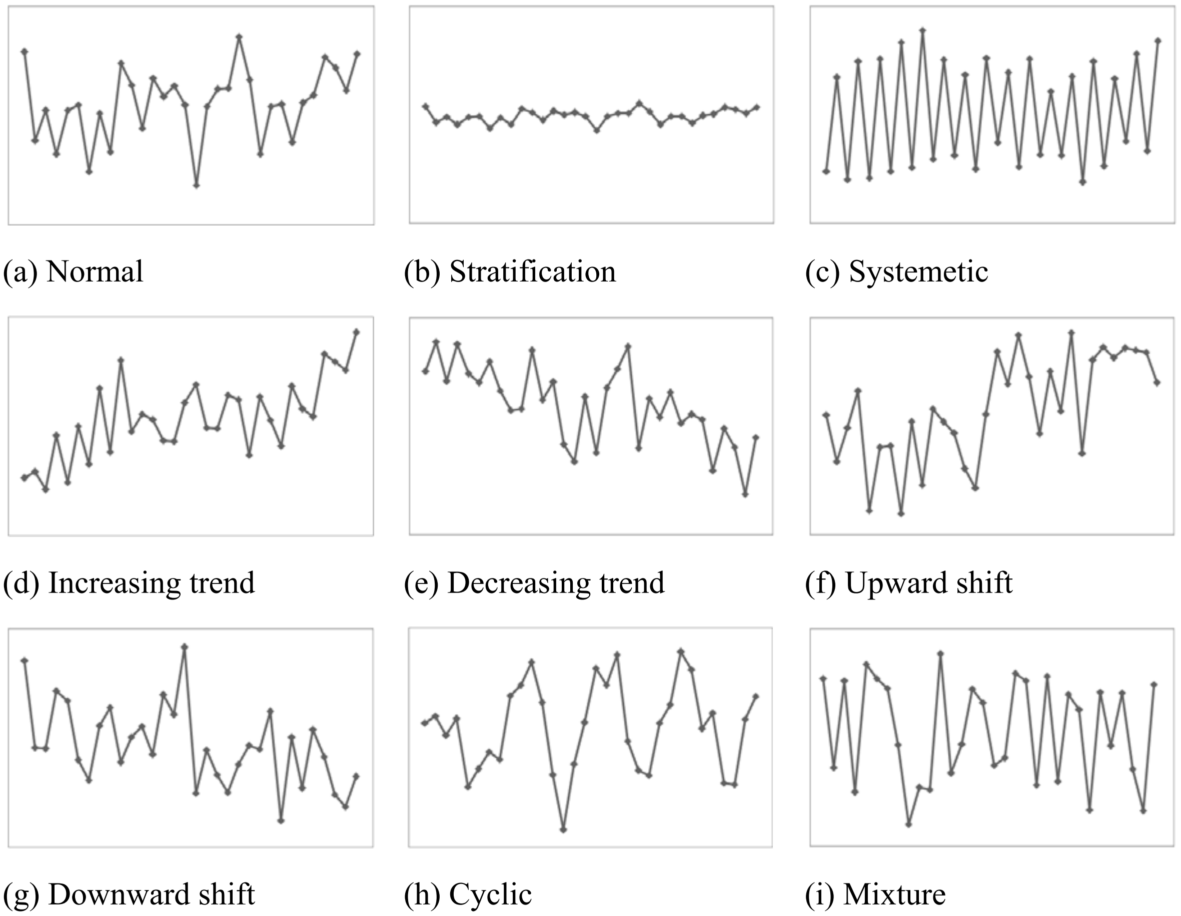
In order to compete in global economy, every manufacturer tries to produce high quality products than their competitors. Statistical process control (SPC) charts are widely used to assess the quality of products being manufactured. Control charts, primarily in the form of $\bar{X}$ chart, are extensively applied to identify the potential process problems in the manufacturing organizations. The $\bar{X}$ chart usually exhibits various types of patterns (Western electric, 1958; Bank, 1989; Montgomery, 2001), e.g. normal (NOR), stratification (STA), systematic (SYS), increasing trend (UT), decreasing trend (DT), upward shift (US), downward shift (DS), cyclic (CYC) and mixture (MIX), as shown in Fig. 1. Only the normal pattern is indicative that the process is under statistical control. The remaining patterns are unnatural and associated with impending problems requiring pre-emptive actions. Identification of the unnatural patterns can greatly minimize the efforts towards troubleshooting and ensure early corrective action(s).

Most of the reported works on control chart pattern (CCP) recognition use raw process data for identifying various CCPs. Evans and Lindsay (1988), Pham and Oztemel (1992a) and Swift and Mize (1995) have developed expert systems that can recognize various CCPs using raw process data as the input. Several techniques have been deployed in the knowledge base design of those expert systems, including template matching and run rules (Nelson, 1984, 1985). One of the main problems with run

* Corresponding author. Tel.: +091-033-2414-6153
E-mail: s_chakraborty00@yahoo.co.in (S. Chakraborty)

© 2012 Growing Science Ltd. All rights reserved.
doi: 10.5267/j.ijiec.2011.09.002

rules is that simultaneous application of all these rules is likely to result in excessive number of false alarms or incorrect recognition. Many other researchers (Pham and Oztemel, 1992b; Hwarng and Hubele, 1993; Cheng, 1997; Guh et al., 1999; Perry et al., 2001; Pacella et al., 2004; Guh and Shiue, 2005) have successfully applied artificial neural networks (ANN) and recognized various CCPs using raw process data. The advantage with neural network is that it is capable of handling noisy measurements requiring no assumption about the statistical distribution of the monitored data. It learns to recognize patterns directly by being presented with the typical example patterns during the training phase. The main disadvantage with neural networks is that the information they contain is implicit and virtually inaccessible to the user. This creates difficulties in understanding how a particular classification decision has been reached and in determining the details of how a given pattern resembles with a particular class. In addition, there is no systematic way to select the topology and architecture of a neural network. In general, this is to be found empirically, which can be time consuming.



(a) Normal (b) Stratification (c) Systemetic

(d) Increasing trend (e) Decreasing trend (f) Upward shift

(g) Downward shift (h) Cyclic (i) Mixture

**Fig. 1.** Various control chart patterns

In recent years, feature-based approaches for CCP recognition have become quite popular. In these approaches, instead of raw process data, extracted features from the raw process data are used as input for CCP recognition. Pham and Wani (1997) and Gauri and Chakraborty (2006, 2009) have utilized different shape features (e.g. slope, number of mean line crossovers etc.), whereas, Hassan et al. (2003) have used statistical features (e.g. mean, skewness etc.) for recognition of various CCPs. The feature-based approaches provide a greater choice of recognition techniques. Pham and Wani (1997) and Gauri and Chakraborty (2006, 2009) have demonstrated that both the properly developed heuristics based on the extracted shape features and an appropriately designed ANN trained using the extracted shape features as input vector representation can efficiently differentiate various CCPs. The feature-based neural network approach reduces the network size and learning time. Since the extracted features represent the main characteristics of the original data in a condensed form, both the feature-based heuristics and neural network approaches can facilitate efficient pattern recognition. The feature-based heuristic approach has a distinct advantage that the practitioners can clearly understand how a particular pattern has been identified by the use of relevant shape features, which is very important in gaining confidence of the practitioners about the usefulness of the CCP recognition system. Pham and Wani (1997) have proposed the feature-based heuristics that can recognize six types of CCPs, e.g. NOR, CYC, UT, DT, US and DS patterns. On the other hand, the feature-based heuristics, as proposed by Gauri and Chakraborty (2006, 2009), can recognize eight types of CCPs, including STA and SYS patterns.

However, as highlighted by Bank (1989) and Montgomery (2001), occurrence of mixture (MIX) pattern is not uncommon in the manufacturing processes. A MIX pattern is indicated when the plotted points tend to fall near or slightly outside the control limits, with relatively few points near the centre line. A MIX pattern is developed by two (or more) overlapping distributions generating the process output. Sometimes, the MIX pattern results from ‘over control’, where the operators make process adjustments too often, responding to random variations in the output rather than systematic causes. A MIX pattern can also occur when the outputs from several sources are fed into a common stream which is then sampled for process monitoring (Montgomery, 2001). Therefore, there is a need for developing feature-based heuristics that can recognize all the nine main types of CCPs, including MIX pattern.

Pham and Wani (1997) have derived the heuristic rules by manually inspecting the feature values in a set of learning samples. However, the manual process of obtaining a good set of heuristics is extremely laborious and thus, it is almost impracticable. Gauri and Chakraborty (2006, 2009) have made use of classification and regression tree (CART) analysis (Breiman et al., 1984) for determining the classification rules. The CART analysis performs binary splits while developing the classification trees. Usage of the tree structured classification algorithm eliminates the burden of manual process and saves time. Loh and Shih (1997) have proposed QUEST (quick unbiased efficient statistical tree) algorithm that also performs binary splits while developing the classification trees. So the QUEST algorithm can be a good alternative to the CART analysis for developing the decision rules in the form of a tree. However, from practitioners’ point of view, it is important to know which algorithm gives the most efficient decision tree with respect to complexity of the tree structure, recognition accuracy and recognition consistency.

In this paper, using the CART-based systematic approach (Gauri and Chakraborty, 2009), a new set of seven most appropriate shape features is selected that can recognize all the nine main types of CCPs, including MIX pattern. Using these selected features, heuristic rules in the form of decision trees are developed using CART as well as QUEST algorithms for recognition of various CCPs. Then, the relative performance of these two algorithms are extensively studied using synthetic pattern data.

## 2. Extraction of selected shape features

Considering a moving observation window of size N = 32 and assuming that a sampling interval in the control chart plot is represented by a linear distance, c = 1σ, Gauri and Chakraborty (2009) have extracted 30 shape features, and then selected a set of seven shape features that can differentiate eight types of CCPs using a CART-based systematic approach.

It can be noted that Gauri and Chakraborty (2009) have extracted various features of CCPs without segmenting as well as by segmenting the available data plot of the observations. It is found that the features extracted without segmenting the observation window are not well capable to discriminate all the nine types of CCPs. Therefore, extraction of features after segmentation of the observation window into a fixed number of segments (pre-defined segments) is considered. The rationale behind it is that if the observations are obtained from an in-control process, the data points in each segment will randomly fluctuate around the mean, otherwise, the patterns formed by those data points in one or more segments will appear to be nonrandom. These differences in pattern characteristics within different segments can be well reflected in the features that may be extracted after segmentation of the observation window. It is found that discrimination between trend and shift patterns still remains difficult. This indicates that there is need to extract additional features that will be capable to discriminate trend and shift patterns.

The basic difference between trend and shift patterns is that in case of trend patterns, the departure of observations from the target value occurs gradually and continuously, whereas, in case of shift patterns, the departure occurs suddenly, and then, the observations hang around the departed value. A shift pattern can be best understood if the time point of occurrence of a shift can be known, and two least square (LS) lines are fitted to the observations before and after the shift. Each LS line will be approximately horizontal to the X-axis for the shift pattern. However, the time point of occurrence of the shift cannot be known exactly. Therefore, a criterion for segmentation of the observation window into two segments is considered. Here, the defined criterion is minimization of the pooled MSE (PMSE) of the LS lines fitted to two segments. It may be noted that, in this segmentation approach, sizes of the two segments may vary in order to satisfy the desired criterion.

Gauri and Chakraborty (2009) have extracted 13 features without segmenting the observation window, 11 features with pre-defined segmentation of the observation window and 6 features with criterion-based segmentation of the observation window. In this paper, the values of all the 30 shape features are first extracted and then using the CART-based systematic approach, a new set of seven features is selected, which can differentiate all the nine main types of CCPs, including MIX pattern. Out of the selected seven features, four features are extracted without segmenting the observation window, two features are extracted after pre-defined segmentation of the observation window and one feature is extracted with criterion-based segmentation of the observation window. These features are described as below:

### *2.1 Features extracted without segmenting the observation window*

a) Sign of slope of the least square (LS) line representing the overall pattern (SB):
The slope (B) of the LS line fitted to the data points in an observation window is given by the following equation:

$$B=\sum_{i=1}^{N} y_i(t_i-\bar{t})/\sum_{i=1}^{N}(t_i-\bar{t})^2 \tag{1}$$

where, $t_i = ic, i = 1,\cdots,N$ is the distance of i$^{\text{th}}$ time point of observation from the origin, c is a constant linear distance used to represent a given sampling interval on the control chart plot, $y_i$ is the observed value of a quality characteristics at i$^{\text{th}}$ time point, N is the size of the observation window and $\bar{t}=\sum_{i=1}^{N} t_i/N$. Then, the feature SB can be defined as follows: SB = 1, if $B \geq 0$ and SB = 0, if $B < 0$. The UT versus DT and US versus DS patterns can be well discriminated using this feature.

b) Ratio between variance of the data points in the observation window (SD$^2$) and mean sum of squares of errors (MSE) of the LS line representing the overall pattern (RVE):
The feature RVE can be extracted using the following expression:

$$\mathrm{RVE}=\left[\frac{1}{N-1}\sum_{i=1}^{N}(y_i-\bar{y})^2\right]\Bigg/\left[\frac{1}{N-2}\left\{\sum_{i=1}^{N}(y_i-\bar{y})^2-\frac{\left(\sum_{i=1}^{N}y_i(t_i-\bar{t})\right)^2}{\sum_{i=1}^{N}(t_i-\bar{t})^2}\right\}\right] \tag{2}$$

The magnitude of RVE for NOR, STA, SYS, CYC and MIX patterns is approximately one, while for trend and shift patterns, it is greater than one.

c) Area between the overall pattern and the LS line per interval in terms of SD$^2$ (ALSPI):

The feature ALSPI can be extracted using the following equation:

$$\text{ALSPI} = [\text{ALS}/(N-1)]/\text{SD}^2\,;\;\; \text{SD}^2 = \textstyle\sum_{i=1}^{N}(y_i - \bar{y})^2/(N-1) \tag{3}$$

where, ALS is the area between the pattern and fitted LS line. The value of ALS can be easily computed by summing the areas of the triangles and trapeziums that are formed by the LS line and overall pattern. The magnitude of ALSPI is the highest for STA pattern, lowest for SYS pattern and intermediate for all other patterns.

d) Proportion of the sum of number of crossovers to mean line and LS line (PSMLSC):

If two successive observations fall in the opposite sides of the mean (centre) line, then only one mean line crossover occurs. This implies that if the mean value is subtracted from the two consecutive observations, the subtracted values will be opposite in sign. Similarly, if the least square estimates of two successive observations fall in the opposite sides of the fitted LS line, then also only one LS line crossover occurs, and the observed value minus the least square estimate for the two successive observations will be opposite in sign. Thus, the feature PSMLSC is extracted using the equation as given below:

$$\text{PSMLSC} = \sum_{i=1}^{N-1}(O_i + O'_i)/2N \tag{4}$$

where, $O_i = 1$ if $(y_i - \bar{y})(y_{i+1} - \bar{y}) < 0$, otherwise, $O_i = 0$ and $\bar{y}$ is the mean value of N data points; $O'_i = 1$ if $(y_i - y'_i)(y_{i+1} - y'_{i+1}) < 0$, otherwise, $O'_i = 0$ and $y'_i$ is the least square estimate of $i^{th}$ data point. The number of mean line as well as LS line crossovers will be maximum for SYS pattern, and minimum for CYC, US and DS patterns. On the other hand, the number of LS line crossovers will be high, but the number of mean line crossovers will be less for UT and DT patterns. For the remaining patterns, both the mean line as well as LS line crossovers will be at intermediate level. Thus, the PSMLSC value is the maximum for SYS pattern, intermediate for NOR, STA, UT, DT and MIX patterns, and lesser for CYC, US and DS patterns.

### *2.2 Features extracted based on segmentation of the observation window*

It is observed that the above four features are not well capable to differentiate between shift and trend patterns and therefore, three more features are selected which are extracted based on segmentation of the observation window. Out of these three features, two are extracted based on pre-determined segmentation (where the segment sizes are fixed) and one is extracted based on criterion-based segmentation (where the segment sizes may vary in order to satisfy the desired criterion).

#### *2.2.1 Pre-determined segmentation*

In this segmentation approach, the total observations are divided into four segments of equal size (N/4). This implies that the window size (N) is so chosen that it can be divisible by 4. The behavior of the process within a segment can be represented by the midpoint of the segment, which is given as

$$\left[\left\{\textstyle\sum_{i=k}^{k+(N/4)-1} t_i/(N/4)\right\}, \left\{\textstyle\sum_{i=k}^{k+(N/4)-1} y_i/(N/4)\right\}\right]$$

where, k = 1, (N/4 + 1), (2N/4 + 1) and (3N/4 + 1) for the first, second, third and fourth segment respectively. A combination of two midpoints can be obtained in $C_2^4 = 6$ ways implying that six straight lines can be drawn passing through the midpoints of these four segments. Similarly, six subsets of N/2 data points can be formed taking a combination of two segments in six ways. So six LS lines can be fitted to six subsets of N/2 data points. The fifth and sixth features described below are extracted based on the properties of straight lines drawn through different combinations of the midpoints and fitted LS lines to different subsets of N/2 data points.

e) Range of slopes of straight lines passing through six pair-wise combinations of midpoints of four equal segments (SRANGE):

$$\text{SRANGE} = \text{maximum}(s_{jk}) - \text{minimum}(s_{jk});\;\; (j = 1,2,3;\;\; k = 2,3,4;\; j < k) \tag{5}$$

In the above equation, $s_{jk}$ represents the slope of the straight line passing through the midpoints of $j^{th}$ and $k^{th}$ segments. The magnitude of SRANGE will be higher for shift pattern than trend pattern. The value of SRANGE will also be higher for CYC pattern than NOR, STA, SYS and MIX patterns, unless each segment of CYC pattern consists of a complete cycle.

f) Ratio of mean sum of squares of errors (MSE) of the LS line fitted to overall data and average MSE of the LS lines fitted to six subsets of $N/2$ data points (REAE)

$$\mathrm{REAE} = \mathrm{MSE}/[\sum_{j,k} \mathrm{MSE}_{jk}/6]\,; \quad (j = 1,2,3;\ k = 2,3,4;\ j < k) \tag{6}$$

where, $\mathrm{MSE}_{jk}$ is the mean sum of squares of errors of the LS line fitted to the observations in $j^{th}$ and $k^{th}$ segments. The magnitude of REAE is greater than one for CYC and shift patterns, and about one for NOR, STA, SYS and trend patterns. In case of MIX pattern, the value of REAE is less than one. The REAE value can differentiate MIX pattern from all other patterns.

*2.2.2 Criterion-based segmentation*

In this segmentation approach, the observation window is divided into two segments based on certain criterion. It may be noted that sizes of the two segments may vary in order to satisfy the desired criterion. Here, the defined criterion is minimization of the pooled MSE (PMSE) of the two LS lines fitted to two segments. Assuming that at least 10 data points are required for fitting a LS line, the LS lines are fitted to all the possible two segments and the segmentation which leads to the minimum PMSE is chosen. The seventh feature described below is extracted based on this criterion-based segmentation of the observation window into two segments.

g) Sum of absolute slope difference between the LS line representing the overall pattern and the individual line segment (SASDPE):

$$\mathrm{SASDPE} = \sum_{j=1}^{2} \left| B - B_j \right| \tag{7}$$

where, B is the absolute slope of the LS line representing the overall pattern and $B_j$ is the slope of the LS line fitted to $j^{th}$ segment. The magnitude of SASDPE is higher for shift patterns than trend patterns. On the other hand, the value of SASDPE will be higher for MIX, CYC and SYS patterns than NOR and STA patterns.

Montgomery and Peck (1982) have highlighted that the prediction based on correlated variables can result in prediction instability and therefore, it is important to ensure that no two selected features are highly correlated. From a set of training samples (see Section 3), all the selected features are extracted and the pair-wise correlation coefficients between them are estimated using the following equation:

$$r_{xy} = \frac{m\sum_{i=1}^{m} x_i y_i - \left(\sum_{i=1}^{m} x_i\right)\left(\sum_{i=1}^{m} y_i\right)}{\left[m\sum_{i=1}^{m} x_i^2 - (\sum_{i=1}^{m} x_i)^2\right]^{1/2}\left[m\sum_{i=1}^{m} y_i^2 - (\sum_{i=1}^{m} y_i)^2\right]^{1/2}} \tag{8}$$

where, $r_{xy}$ is the correlation coefficient between two features represented by x and y, and m is the number of pairs of values of the two features. Since a set of training samples contains 9000 pattern data, here the value of m is 9000. Table 1 shows the values of pair-wise correlation coefficients for the seven selected features computed from the set of training samples. This table reveals that the degree of association between the selected shape features is considerably low. The highest value of pair-wise correlation coefficient is 0.58 only. Therefore, the selected set of features is considered to be appropriate for developing the decision trees for CCP recognition.

**Table 1**
Pair-wise correlation coefficients for the selected features

| Feature | SB | RVE | ALSPI | PSMLSC | SRANGE | REAE | SASDPE |
|---|---|---|---|---|---|---|---|
| SB | 1.00 | 0.01 | -0.01 | 0.23 | -0.16 | -0.09 | -0.05 |
| RVE | | 1.00 | -0.26 | -0.34 | 0.03 | 0.18 | 0.01 |
| ALSPI | | | 1.00 | -0.04 | -0.34 | -0.05 | -0.41 |
| PSMLSC | | | | 1.00 | -0.43 | -0.36 | -0.19 |
| SRANGE | | | | | | 0.58 | 0.36 |
| REAE | | | | | | 1.00 | 0.13 |
| SASDPE | | | | | | | 1.00 |

## 3. Generation of sample patterns

Ideally, sample patterns should be collected from a real manufacturing process. Since, a large number of patterns is required for developing and validating a CCP recognizer, and as those are not economically available from the manufacturing processes, simulated data are often used. A large window size can lower the recognition efficiency by increasing the time required to detect the patterns. On the other hand, the patterns will not develop appropriately if the window size is too small, and thus, discrimination of various patterns will become difficult. In this paper, therefore, a window with 32 observations is considered, which is neither too large nor too small. The equations along with the corresponding pattern parameters used for simulating the nine basic CCPs are given in Table 2. In this table, i (i = 1,2,3,…,32) is the discrete time point at which the pattern is sampled, $r_i$ is random value of a standard normal variate at $i^{th}$ time point and $y_i$ is the sample value at $i^{th}$ time point.

The values of different parameters for the unnatural patterns are randomly varied in a uniform manner between the limits shown. A set of 9000 (1000×9) sample patterns is generated from 1000 series of standard normal variate. Multiple sets of learning samples as well as test samples are required to rigorously evaluate the recognition and generalization performance of the heuristic-based CCP recognizer that can be developed based on the selected set of shape features. In this paper, ten sets of learning and ten sets of test samples of size 9000 each are generated for the purpose of experimentation. The only difference between these 20 sets of sample patterns is in the random generation of standard normal variate and in the values of different pattern parameters within their respective limits.

**Table 2**
Parameters for simulating control chart patterns

| Pattern | Pattern parameters | Parameter values | Pattern equation |
|---|---|---|---|
| NOR | • Mean (μ)<br>• Standard deviation (σ) | 80<br>5 | $y_i = \mu + r_i\sigma$ |
| STA | • Random noise (σ′) | 0.2σ to 0.4σ | $y_i = \mu + r_i\sigma'$ |
| SYS | • Systematic departure (d) | 1σ to 3σ | $y_i = \mu + r_i\sigma + d \times (-1)^i$ |
| UT | • Gradient (g) | 0.05σ to 0.1σ | $y_i = \mu + r_i\sigma + ig$ |
| DT | • Gradient (g) | −0.1σ to −0.05σ | $y_i = \mu + r_i\sigma + ig$ |
| US | • Shift magnitude (s)<br>• Shift position (P) | 1.5σ to 2.5σ<br>9, 17, 25 | $y_i = \mu + r_i\sigma + ks;$<br>$k = 1$ if $i \ge P$, else $k = 0$ |
| DS | • Shift magnitude (s)<br>• Shift position (P) | −2.5σ to −1.5σ<br>9, 17, 25 | $y_i = \mu + r_i\sigma + ks;$<br>$k = 1$ if $i \ge P$, else $k = 0$ |
| CYC | • Amplitude (a)<br>• Period (T) | 1.5σ to 2.5σ<br>8 and 16 | $y_i = \mu + r_i\sigma + a\sin(2\pi i/T)$ |
| MIX | • Process mean (m)<br>• A random number (p) | 1.5σ to 2.5σ<br>0 to 1 | $y_i = \mu + r_i\sigma + (-1)^w m$<br>$w = 0$ if $p < 0.4$, $w = 1$ if $p \ge 0.4$ |

## 4. Developing decision trees for pattern recognition

The developed feature-based decision trees (heuristics) use simple IF... (condition)…THEN…(action)…heuristic rules. The conditions for the rules set the threshold values of the features and the actions are the classification decisions. The set of heuristics, arranged as a decision tree, can provide easily understood and interpreted information regarding the predictive structure of the data for various features. Given a set of extracted features from the learning samples, classification tree programs, like CART and QUEST algorithms can generate a set of heuristics based on the features arranged as binary decision tree, which can be used for recognition of control chart patterns. Both these tree-structured classification algorithms allow automatic selection of the 'right-sized' tree that has the optimal predictive accuracy. The procedures for the 'right-sized' tree selection are not foolproof, but at least, they take the subjective judgment out of the process of choosing the 'right-sized' tree and thus avoid 'over fitting' and 'under fitting' of the data.

The CART performs an exhaustive grid search of all the possible univariate splits while developing a classification tree. The search procedure for CART algorithm (Breiman et al., 1984) can be lengthy when there are a large number of predictor variables with many levels and it is biased towards choosing the predictor variables with more levels for splits (Loh and Shih, 1997). On the other hand, QUEST algorithm (Loh and Shih, 1997) employs modification of the recursive quadratic discriminant analysis and includes a number of innovative features for improving the reliability and efficiency of the classification tree that it computes. The QUEST algorithm is fast and unbiased. Its lack of bias in variable selection for splits is a distinct advantage when some predictor variables have few levels and other have many. Moreover, QUEST does not sacrifice the predictive accuracy for speed (Lim et al., 1997). Therefore, it is planned to adopt both the CART and QUEST (available in STATISTICA package) algorithms for determining the tree-structured classification rules for CCP recognition and evaluate their relative performance.

Extracting the selected shape features from a set of training samples consisting of all the nine types of CCPs and subjecting the feature values along with the pattern identification codes to CART or QUEST analysis in STATISTICA produce a pattern classification tree, i.e. feature-based heuristic rules for pattern classification. However, the performance of these rules must be evaluated before implementing them for the purpose of pattern classification in real life situations. For this, verification samples containing different types of CCPs are required. For rigorous evaluation of CART or QUEST-based heuristic rules, multiple sets of training and verification samples are necessary. As those are not economically available from the manufacturing processes, simulated data are often used.

Development of the classification tree based on CART or QUEST algorithm from the data set containing extracted values of seven features and pattern identification codes using STATISTICA software requires the following steps:

*Step 1*: Open the data file containing the extracted values of the seven features and pattern identification codes.

*Step 2*: From the statistical analysis module, called 'Statistics', select the 'Classification tree' analysis module under the 'Multivariate exploratory techniques'.

*Step 3*: Click on the 'Quick' or 'Advanced' tab and then select the dependent variables, categorical predictor variables and ordered predictor variables. In this case, pattern identification code is selected as the dependent variable, features having discrete values, e.g. sign of slope are selected as the categorical predictor variables and other features having continuous values are selected as the ordered predictor variables.

*Step 4*: Select the method to be used for developing the classification tree. Selection of 'C&RT style exhaustive search for univariate splits' implies CART algorithm. On the other hand, 'Discriminant-based univariate splits for categorical and ordered predictor' selects QUEST algorithm for use.

*Step 5*: Specify the prior probabilities for different patterns and the corresponding misclassification cost. For application of CART algorithm, it is required to specify the measure of 'Goodness of fit for a split' also.
*Step 6*: Click on 'Stopping options', and then specify 'Stopping rule' and 'Stopping parameters'.
*Step 7*: Click on 'OK' to run the selected algorithm. It results in the classification tree along with the other details.

### *4.1. Experimentation*

Ten different sets of training samples of size 9000 each and ten different sets of test samples of size 9000 each are generated. For generation of a set of training or verification sample, 1000 series of standard normal data of size 32 each are generated first. Using these standard normal data, 1000 series of pattern data are created for each pattern class. Thus, each set of training or verification sample contains 9000 (1000 x 9) sample patterns and each set contains equal number of pattern data for each pattern class. For developing a classification tree from a set of training samples, the extracted values of the seven features along with the pattern identification codes are subjected to classification tree analysis with the following specifications:

a) prior probabilities for different patterns – proportional to class size

b) misclassification cost of a pattern – equal for all the patterns

c) measure of goodness of fit for a split – Gini index (for application of CART algorithm).

The 'stopping rule' and 'stopping parameters' for application of CART algorithm are specified as follows:

a) stopping rule – prune on misclassification error

b) stopping parameters – i) value of 'n' for 'Minimum n' rule = 1, and ii) value of 'δ' for 'δ standard error' rule = 0.1.

On the other hand, the 'stopping rule' and 'stopping parameters' for application of QUEST algorithm are set as below:

a) stopping rule – prune on misclassification error

b) stopping parameters – i) value of 'n' for 'Minimum n' rule = 5, and ii) value of 'δ' for 'δ standard error' rule = 1.0.

Ten different decision trees, i.e. ten different heuristic-based CCP recognizers are obtained by subjecting the extracted feature values from the ten sets of training samples to classification tree analysis using CART algorithm. These ten heuristic-based CCP recognizers are labelled as 1.01-1.10 in Table 3. Similarly, ten different heuristic-based CCP recognizers are also obtained by subjecting the extracted feature values from the ten sets of training samples to classification tree analysis using QUEST algorithm. These heuristic-based CCP recognizers are labelled as 2.01-2.10 in Table 4. The CCP recognition performance of each heuristic-based recognizer is then evaluated using all the ten sets of test samples.

## 5. Results and discussions

The training (learning) samples are required to develop the pattern classification rules and verification (test) samples are needed to evaluate the general ability of the developed classification rules to correctly recognize various control chart patterns. Multiple sets of learning as well as test samples are necessary to rigorously evaluate the recognition and generalization performance of the developed heuristic rules. So any work aiming to CCP recognition system must include both the training and recall phases. In training phase, the heuristic rules are developed and in recall phase, the general performance of the developed rules is evaluated. The training and recall (verification) performance of the ten CART-based and ten QUEST-based CCP recognizers are shown in Table 3 and 4 respectively. It is observed that the overall mean percentage of correct recognition achieved by the CART-based recognizers at the training and recall phases (95.53% and 94.67% respectively) are higher than those obtained by the QUEST-based recognizers (93.44% and 92.76% respectively). It may be noted that the recognition performance of both the CART-based and QUEST-based recognizers at the recall phase is inferior to that achieved during the training phase. The percentage of correct recognition for the CART and QUEST-based recognizers at the recall phase ranges from 93.18% to 96.08% and from 91.81% to 93.85% respectively. Statistical test is required to rule out that the difference in recognition performance between the CART and QUEST-based recognizers is not attributable to the sampling fluctuations. In this paper, since all the ten sets of verification samples are subjected to classification using the CART and QUEST-based recognizers derived from the same set of training (learning) samples, the most appropriate statistical test for the difference in recognition performance is the paired t-test. So, paired t-tests at 1% significance level (i.e. $\alpha = 0.01$) are conducted for 10 pairs of CART and QUEST-based recognizers for their performance in terms of percentage of correct classification. The results of statistical significance tests are summarised in Table 5. These results suggest that the difference in recognition accuracy between these two types of recognizers is significant. This confirms that the CART-based recognizers give better recognition performance compared to the QUEST-based recognizers.

On the other hand, comparison of the standard deviation (SD) value of recognition performance of the ten CART-based and ten QUEST-based recognizers at the recall phase indicates that the variation in recognition performance is consistently lesser for the QUEST-based recognizers than the CART-based recognizers. This is indicative that the recognition performance of the QUEST-based recognizers is more consistent than the CART-based recognizers. However, the decision trees derived using the QUEST algorithm are found to be more complex. The average number of terminal nodes in the CART-based decision trees is observed to be 20.9, whereas, for the QUEST-based decision trees, the average number of terminal nodes is found as 28.9. The best recognizer with respect to the recognition performance is recognizer number 1.05 (CART-based) and it is shown in Fig. 2. On the other hand, recognizer number 2.08 (QUEST-based), as shown in Fig. 3, is the best one with respect to recognition consistency.

**Table 3**
Learning and recall performance of the CART-based recognizers

| Recognizer number | Learning phase | | Recall phase | |
|---|---|---|---|---|
| | Number of terminal nodes | Percentage of correct classification | Percentage of correct classification | |
| | | | Mean | Standard deviation |
| 1.01 | 22 | 95.47 | 95.22 | 1.65 |
| 1.02 | 20 | 94.16 | 93.57 | 1.78 |
| 1.03 | 23 | 95.33 | 93.25 | 1.73 |
| 1.04 | 22 | 95.84 | 95.27 | 1.82 |
| **1.05** | **18** | **96.16** | **96.08** | **1.99** |
| 1.06 | 19 | 95.09 | 93.93 | 1.67 |
| 1.07 | 21 | 96.00 | 95.86 | 2.06 |
| 1.08 | 23 | 95.58 | 93.18 | 1.88 |
| 1.09 | 21 | 96.02 | 95.73 | 1.77 |
| 1.10 | 20 | 95.64 | 94.60 | 1.89 |
| Mean | 20.90 | 95.53 | 94.67 | - |

**Table 4**
Learning and recall performance of the QUEST-based recognizers

| Recognizer number | Learning phase | | Recall phase | |
|---|---|---|---|---|
| | Number of terminal nodes | Percentage of correct classification | Percentage of correct classification | |
| | | | Mean | Standard deviation |
| 2.01 | 31 | 91.33 | 92.76 | 1.05 |
| 2.02 | 28 | 93.24 | 92.58 | 1.11 |
| 2.03 | 28 | 92.84 | 91.81 | 1.03 |
| 2.04 | 27 | 94.16 | 92.97 | 1.13 |
| 2.05 | 30 | 94.38 | 93.59 | 1.26 |
| 2.06 | 28 | 94.56 | 92.42 | 1.07 |
| 2.07 | 31 | 92.29 | 92.08 | 1.17 |
| **2.08** | **27** | **93.87** | **91.94** | **0.98** |
| 2.09 | 31 | 94.04 | 93.61 | 1.13 |
| 2.10 | 28 | 93.66 | 93.85 | 1.17 |
| Mean | 28.90 | 93.44 | 92.76 | - |

**Table 5**
Statistical significance test for difference in recall performance in the CART and QUEST-based CCP recognizers

| Hypothesis | $t_{statistics}$ | $t_{critical}$ | Decision |
|---|---|---|---|
| $H_0: \mu_{CART} - \mu_{QUEST} = 0$<br>$H_1: \mu_{CART} - \mu_{QUEST} > 0$ | 6.67 | 2.82 | Reject $H_0$ |

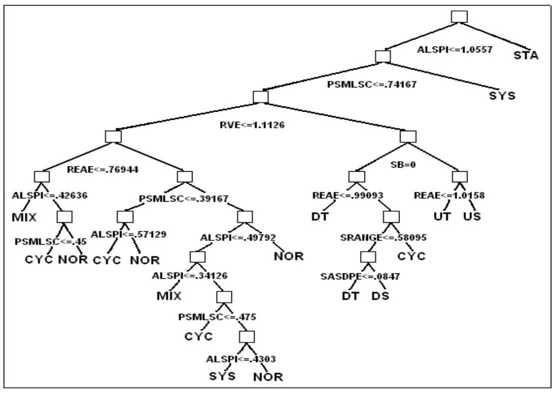


**Fig. 2.** Decision tree for CCP recognition derived using CART algorithm

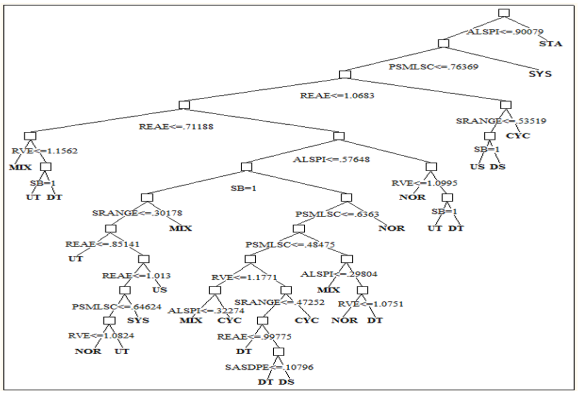


**Fig. 3.** Decision tree for CCP recognition obtained using QUEST algorithm

## 6. Conclusions

One important advantage of the feature-based decision trees for CCP recognition is that in this approach, the practitioners can clearly understand how a particular pattern has been identified by the use of relevant shape features. The feature-based decision trees, reported in the literature, can recognize eight types of CCPs using extracted values of seven shape features. In this paper, decision trees are developed using CART and QUEST algorithms, based on a different set of seven most useful shape features, which can recognize nine main CCPs, including mixture pattern. Relative performance of the CART-based and QUEST-based decision trees is extensively studied using simulated pattern data. The results show that the CART-based decision trees result in better recognition performance but lesser consistency, whereas, the QUEST-based decision trees give better consistency but lesser recognition performance.

## References

Bank, J. (1989). *Principles of Quality Control.* Singapore: John Wiley & Sons.

Breiman, L., Friedman, J.H., Olshen, R.A., & Stone, C.J. (1984). *Classification and Regression Trees.* Monterey, CA: Wadsworth and Brooks.

Cheng, C.S. (1997). A neural network approach for the analysis of control chart patterns. *International Journal of Production Research,* 35, 667-697.

Evans, J.R., & Lindsay, W.M. (1988). A framework for expert system development in statistical quality control. *Computers and Industrial Engineering*, 14, 335-343.

Gauri, S.K., & Chakraborty, S. (2006). Feature-based recognition of control chart patterns. *Computers and Industrial Engineering,* 51, 726-742.

Gauri, S.K., & Chakraborty, S. (2009). Recognition of control chart patterns using improved selection of features. *Computers and Industrial Engineering,* 56, 1577-1588.

Guh, R.S., Zorriassatine, F., Tannock, J.D.T., & O'Brien, C. (1999). On-line control chart pattern detection and discrimination - A neural network approach. *Artificial Intelligence in Engineering,* 13, 413-425.

Guh, R.S., & Shiue, Y.R. (2005). On-line identification of control chart patterns using self-organizing approaches. *International Journal of Production Research,* 43**,** 1225-1254.

Hassan, A., Nabi Baksh, M.S., Shaharoun, A.M., & Jamaluddin, H. (2003). Improved SPC chart pattern recognition using statistical features. *International Journal of Production Research,* 41, 1587-1603.

Hwarng, H.B., & Hubele, N.F. (1993). Back-propagation pattern recognizers for $\bar{\mathrm{X}}$ control charts: Methodology and performance. *Computers and Industrial Engineering,* 24, 219-235.

Lim, T.S., Loh, W.Y., & Shih, Y.S. (1997). *An empirical comparison of decision trees and other classification methods*. Madison: University of Wisconsin, Department of Statistics, Technical Report 979.

Loh, W.Y., & Shih, Y.S. (1997). Split selection methods for classification trees. *Statistica Sinica,* **7**, 815-840.

Montgomery, D.C., & Peck, E.A. (1982). *Introduction to linear regression analysis.* New York: John Wiley & Sons.

Montgomery, D.C. (2001). *Introduction to statistical quality control.* New York: John Wiley & Sons.

Nelson, L.S. (1984). The Shewhart control chart - Test for special causes. *Journal of Quality Technology,* 16, 237-239.

Nelson, L.S. (1985). Interpreting Shewhart $\bar{X}$ control chart. *Journal of Quality Technology,* 17, 114-117.

Pacella, M., Semeraro, Q., & Anglani, A. (2004). Manufacturing quality control by means of a fuzzy ART network trained on natural process data. *Engineering Applications of Artificial Intelligence,* 17, 83-96.

Perry, M.B., Spoerre, J.K., & Velasco, T. (2001). Control chart pattern recognition using back propagation artificial neural networks. *International Journal of Production Research,* 39, 3399-3418.

Pham, D.T., & Wani, M.A. (1997). Feature-based control chart pattern recognition. *International Journal of Production Research,* 35, 1875-1890.

Pham, D.T., & Oztemel, E. (1992a). XPC: An on-line expert system for statistical process control. *International Journal of Production Research,* 30, 2857-2872.

Pham, D.T., & Oztemel, E. (1992b). Control chart pattern recognition using neural networks. *Journal of System Engineering,* 2, 256-262.

Swift, J.A., & Mize, J.H. (1995). Out-of-control pattern recognition and analysis for quality control charts using LISP-based systems. *Computers and Industrial Engineering,* 28, 81-91.

Western Electric (1958). *Statistical Quality Control Handbook.* Indianapolis: Western Electric Company.