Revista de Matemática: Teoría y Aplicaciones 2009 **16**(1) : 30–42
cimpa – ucr issn: 1409-2433

Árboles de clasificación para el análisis de gráficos de control multivariantes

Matías Gámez Martínez* Esteban Alfaro Cortés †
José Luis Alfaro Navarro‡ Noelia García Rubio.§

Recibido/Received: 20 Feb 2008 — Aceptado/Accepted: 25 Jul 2008

Resumen

En control estadístico de la calidad, una de las herramientas más utilizadas son los gráficos de control. El principal problema de los gráficos de control multivariantes radica en que sólo indican que se ha producido un cambio en el proceso, pero no dice cuál o cuáles de las variables son las que originan este cambio. En la literatura especializada existen muchas aproximaciones para solucionar este problema, si bien, la más utilizada consiste en la descomposición del estadístico T^2. En este trabajo se propone un método alternativo mediante la aplicación de árboles de clasificación. Los resultados obtenidos muestran que estos árboles de clasificación constituyen una buena herramienta para completar la interpretación de los gráficos de control multivariantes.

Palabras clave: Control estadístico de la calidad, T^2 de Hotelling, Árboles de clasificación.

Abstract

In statistical quality control, one of the most widely used tools are the control charts. The main problem of the multivariate control charts lies in that they only indicate that a change in the process has happened, but they do not show which variable or variables are the source of this change. In the specialized literature there are many approaches to tackle this problem, although the most usual consists on the decomposition of the T^2 statistic. In this research, we propose an alternative method through the application of classification trees. The results show that this method constitutes a good tool to help to interpret the multivariate control charts.

*Facultad de Ciencias Económicas y Empresariales de Albacete, Universidad de Castilla-La Mancha. E-mail:Matias.Gamez@uclm.es

†Misma dirección que M. Gámez. E-mail: Esteban.Alfaro@uclm.es

‡Misma dirección que M. Gámez. E-mail: JoseLuis.Alfaro@uclm.es

§Misma dirección que M. Gámez. E-mail: Noelia.Garcia@uclm.es

Keywords: Statistic Process Control, T^2 Hotelling, Classification trees.

Mathematics Subject Classification: 62H15, 68T01, 62H30,62M45, 62P30

1 Introducción

En control estadístico de procesos (SPC), las técnicas univariantes están diseñadas para controlar la calidad mediante el análisis de una única característica de calidad. Sin embargo, en los procesos de producción actuales, suelen existir varias características que influyen de forma conjunta e interrelacionada en la calidad final de los productos. Una solución consiste en controlar todas de forma simultánea mediante técnicas de control multivariantes. De esta forma no sólo vamos a analizar el efecto de cada una de las características sobre la calidad, sino que vamos a considerar también el efecto de las interacciones entre ellas.

Dentro de estas técnicas multivariantes destacan tres: T^2 de Hotelling, MEWMA y MCUSUM. En este trabajo nos vamos a centrar en el gráfico de control T^2 de Hotelling, extensión multivariante del gráfico Shewhart y, sin duda, el gráfico multivariante más utilizado. En la mayoría de los métodos de control multivariantes se recurre a la determinación de cierto estadístico que resuma la información, en este caso el estadístico T^2. Este estadístico es un escalar que combina información de medidas de dispersión ($\mathbf{\Sigma}$) y posición (μ) de las variables y que, para observaciones individuales, se va a determinar como:

$$T^2 = (\mathbf{x} - \mu)^{\mathbf{t}} \mathbf{\Sigma}^{-\mathbf{1}} (\mathbf{x} - \mu) \tag{1}$$

En este caso, si se conocen los parámetros poblacionales, la distribución del estadístico T^2 va a ser una chi-cuadrado[1] y, por contra, si no se conocen los parámetros poblacionales habrá que recurrir a una estimación de los parámetros ($\mathbf{S}$ y $\bar{\mathbf{x}}$, respectivamente) necesarios para establecer el gráfico de control mediante la expresión:

$$T^2 = (\mathbf{x} - \bar{\mathbf{x}})^{\mathbf{t}} \mathbf{S}^{-\mathbf{1}} (\mathbf{x} - \bar{\mathbf{x}}) \tag{2}$$

Para conocer la distribución de este estadístico es necesario distinguir dos situaciones: en el primer caso, si el vector de observaciones es independiente de las estimaciones de los parámetros, es decir, no se usa en la estimación de éstos, el límite superior de control va a venir dado por:

$$LSC = \frac{p\ (n+1)(n-1)}{n^2 - np} F_{\alpha;p,n-p} \tag{3}$$

donde $F_{\alpha;p,n-p}$ es el $100 \cdot (1-\alpha)$ cuantil de la distribución F de Snedecor con p y $n-p$ grados de libertad.

O bien, cuando toda la información que contiene el vector de observaciones es usada en la estimación de los parámetros, en cuyo caso el límite de control va a ser:

[1] Las propiedades distribucionales del estadístico T^2 han sido ampliamente analizadas por Tracy, Young et al. (1992) y Mason & Young (2002).

$$LSC = \frac{(n-1)^2}{n}\beta_{\alpha;p/2,(n-p-1)/2}$$

donde $\beta_{\alpha;p/2,(n-p-1)/2}$ es el $100\cdot(1-\alpha)$ cuantil de la distribución beta con p y $n-p$ grados de libertad. Vamos a considerar la distribución dada en la ecuación (3) con un valor de α (probabilidad de error tipo I) de 0.05.

Con las técnicas gráficas de control multivariante, detectar una situación fuera de control es relativamente fácil, ya que el análisis es similar al caso univariante, pero determinar las causas que han provocado ese cambio es más complicado. En este trabajo proponemos la aplicación de una técnica de clasificación para determinar la variable o variables que han causado la situación fuera de control, en concreto, la aplicación de árboles de clasificación como alternativa a las redes neuronales artificiales que están siendo aplicadas al problema recientemente.

2 Métodos para interpretar señales fuera de control en control estadístico de la calidad

El problema de la interpretación de señales fuera de control en los gráficos multivariantes ha frenado el desarrollo de estas técnicas en la industria. El gráfico T^2 de Hotelling no indica qué variable o variables, de las que se están midiendo, han provocado la situación fuera de control. Esto requiere un trabajo posterior nada simple para poder encontrar las variables que han cambiado, dado que una situación fuera de control puede deberse a una o varias variables fuera de control o a un cambio en la relación existente entre las variables.

Se han desarrollado algunas técnicas para ayudar en la interpretación de las señales fuera de control, si bien la más utilizada consiste en analizar gráficos de control univariantes para cada una de las características de calidad. Este camino presenta ciertos inconvenientes: el primero es que cuando hay muchas variables, esta técnica puede resultar tediosa por el gran número de gráficos univariantes para analizar; y el segundo es que normalmente una señal fuera de control no es causada sólo por una variable, sino más bien por la relación existente entre variables, por lo que en muchas situaciones los gráficos univariantes no van a mostrar señales fuera de control.

Para facilitar la interpretación de señales fuera de control en gráficos multivariantes hay distintas propuestas en la literatura especializada entre las que destacan: Alt (1985) desarrolló gráficos de medias individuales con un control tipo Bonferroni; Hayter & Tsui (1994) utilizaron un procedimiento de intervalos de confianza simultáneos tipo Bonferroni para cada una de las características de calidad. Para detectar la variable que ha causado el cambio también se puede recurrir a la descomposición del estadístico T^2 de forma que nos mida la influencia de cada una de las variables. Si T^2 es el valor del estadístico y $T^2_{(i)}$ es el valor de ese estadístico para todas las variables del proceso excepto la i-ésima, podemos calcular un indicador de la contribución de la variable i-ésima sobre el conjunto de la siguiente forma:

$$d_i = T^2 - T^2_{(i)}$$

Cuando aparece una situación fuera de control en un gráfico de control multivariante es conveniente calcular esta contribución para cada una de las variables y centrar nuestra atención en aquellas variables cuya contribución sea superior.

Siguiendo otros enfoques, Jackson (1980) propone representar componentes principales en lugar de las variables originales, con el posible problema de interpretación de dichas componentes; Murphy (1987) recomienda un método que consiste en desarrollar un procedimiento basado en un análisis discriminante que permite agrupar las observaciones y Mason et al. (1995) proponen una descomposición del estadístico T^2 de una forma más compleja que la anterior. Esta descomposición del estadístico T^2 aparece recogida en Mason & Young (2002) y consiste en dividir el estadístico en dos componentes: T^2 no condicionada, que recoge el valor del estadístico T^2 para cada una de las variables y T^2 condicionada, que recoge el comportamiento de dicho estadístico basado en los residuos de la regresión de cada variable sobre las demás. De esta forma se pretende recoger las variaciones que se producen en la estructura de correlación de las variables.

Otra alternativa, más reciente, en la interpretación de gráficos de control multivariantes, consiste en el uso de redes neuronales artificiales que permiten automatizar la selección de las variables que han provocado el cambio en el proceso, de modo que en el momento que el gráfico T^2 detecta una salida de control, la red neuronal identificaría estas variables. Algunos ejemplos de aplicación de redes neuronales en control estadístico de la calidad son: Cheng (1995, 1997), Chang (1996), Zorriassatine (1998), Guh & Tannock (1999), Guh (2003), Noorosana, et al. (2003), Niaki & Abassi (2005), Aparisi et al. (2006) y Guh (2007).

Por lo tanto, parece conveniente utilizar el gráfico de control T^2 de Hotelling no de forma aislada, sino complementado con alguna técnica de análisis de señales fuera de control, obteniendo así una interpretación más clara de los resultados obtenidos.

En este trabajo proponemos la aplicación de árboles de clasificación para poder determinar la variable o variables que han causado la situación fuera de control. En este sentido, los árboles de clasificación son una técnica alternativa a las redes neuronales que presentan un buen comportamiento a la hora de interpretar las señales fuera de control puestas de manifiesto por el gráfico. A continuación describimos brevemente estos árboles de clasificación.

3 Árboles de clasificación

Los árboles de clasificación (Breiman et al., 1984) pueden verse como un método de clasificación no paramétrico, lo que nos permite flexibilidad, adems de ser capaces de trabajar de forma eficiente con variables cualitativas. Un árbol de clasificación se representa gráficamente mediante nodos y ramas. Cada nodo simboliza una cuestión o decisión sobre una de las características de los ejemplos. El nodo inicial se suele llamar nodo raíz. De cada nodo puede salir dos o más ramas, dependiendo de que la respuesta a la cuestión planteada sea binaria o no. Finalmente, se alcanzan los nodos terminales u hojas y se toma

una decisión sobre la clase a asignar. Cuando se presente un nuevo ejemplo o patrón al árbol, éste lo filtrará a lo largo de los test que contienen los nodos. Cada test tiene salidas mutuamente excluyentes y exhaustivas, lo que significa que los ejemplos que se asignen a una de las salidas, no se pueden asignar a otra y además todos los ejemplos se asignarán a una de las salidas. Es decir, ningún ejemplo se asignará a dos salidas de un mismo test, pero tampoco habrá ningún ejemplo que no se asigne a ninguna salida.

Los árboles de clasificación pueden trabajar tanto con variables continuas como con variables categóricas de dos o más modalidades. A la hora de construir un árbol de clasificación para un conjunto de entrenamiento concreto son muchas las posibilidades que existen, por lo que es inabordable examinarlas una a una y, por tanto, se hace necesario el establecimiento de una serie de mecanismos para buscar alguno de forma óptima.

La construcción del árbol se realiza durante la fase de aprendizaje, que puede esquematizarse en los siguientes pasos que se repiten recursivamente:
1. Cada nodo se parte en función de una prueba que se plantea sobre el valor de alguna de las características que describen los ejemplos. En el caso binario, verdadero o falso, los ejemplos que cumplen la condición se asignarán a uno de los nodos hijos y los restantes, al otro. Salvo el nodo raíz, cuando se parte un nodo, éste pasa a ser un nodo interno o intermedio. De entre todas las posibles preguntas y particiones que se pueden hacer, habrá que elegir aquéllas que lleven a obtener un mejor resultado, es decir, que obtenga un mayor incremento en la homogeneidad o pureza de los nodos hijos con respecto al nodo padre. Para ello, hay que establecer una medida de impureza y entre los criterios más utilizados se encuentran la medida de Entropía, que utilizan algoritmos tales como el ID3 (Quinlan, 1986) y el C4.5 (Quinlan, 1993) y la basada en el Índice de Gini utilizada por el algoritmo CART (Breiman et al., 1984). La medida de Entropía en un nodo t, se calcula como

$$i(t) = -\sum_{j=1}^{q} p(j|t) \cdot \ln\left(p(j|t)\right) \tag{4}$$

donde se asume que $0 \cdot ln(0) = 0$ y se estima $p(j|t)$ como la proporción de ejemplos de la clase j en el nodo t, donde $j = 1, 2, \ldots, q$, es decir

$$p(j|t) = \frac{N_j(t)}{N(t)}$$

donde $N(t)$ y $N_j(t)$ son el número de ejemplos total del nodo t y el número de ejemplos de la clase j en t, respectivamente.

El Índice de Gini mide la concentración u homogeneidad de las clases en un nodo y se calcula como

$$i(t) = -\sum_{\substack{i,j=1 \\ i \neq j}}^{q} p(i|t) \cdot p(j|t) = 1 - \sum_{j=1}^{q} \left(p(i|j)\right)^2 \tag{5}$$

2. La condición de parada detiene el proceso de partición de nodos, cuando un nodo cumple esta condición se dice que es un nodo terminal u hoja. Existen varias posibilidades, entre

otras, detener la división de los nodos cuando sean puros (todos los ejemplos del nodo hoja son de la misma clase) o cuando su tamaño sea inferior a un determinado umbral, o superen un determinado nivel de pureza, considerada en términos de la proporción de la clase mayoritaria, o utilizar la ganancia de información o reducción de impureza como criterio para detener el crecimiento del árbol. Los ejemplos pertenecientes a un nodo hoja tendrán cierta homogeneidad, por lo que al nodo se le asigna una etiqueta con la clase mayoritaria.

Una alternativa a la detención del desarrollo del árbol para evitar el sobreajuste de éste son los métodos que desarrollan el árbol hasta que los nodos hoja sean puros o muy pequeños, para después aplicar algún mecanismo de poda al árbol completo. A priori, el primer método puede parecer más rápido y mejor; sin embargo, Breiman et al. (1984) defienden la poda a posteriori frente a la interrupción del crecimiento. Según ellos resulta más eficiente la poda ya que permite que uno o varios subárboles de un nodo permanezcan y el resto desaparezcan, mientras que si se interrumpe el crecimiento, todas las ramas son eliminadas.

4 Detalles del método propuesto

El método propuesto en este trabajo se realiza en dos etapas. En primer lugar se aplica el gráfico de control T^2 de Hotelling para detectar la aparición de observaciones fuera de control. Cuando se produce una señal fuera de control se pasa a la siguiente etapa, que consiste en utilizar los árboles de clasificación para determinar la variable o variables que han producido dicha alteración.

Para mostrar cómo funciona este procedimiento, hemos considerado conveniente trabajar con ejemplos que hayan demostrado previamente su utilidad para esta labor. El trabajo reciente de Niaki & Abbasi (2005) se apoya en tres ejemplos que abarcan distintos niveles de dificultad al trabajar con dos, tres y cuatro variables. Además, tendremos la posibilidad de comparar los resultados, siempre con las debidas cautelas y sólo a efectos referenciales.

Ejemplo 1: El caso de dos variables.

Comenzamos con un ejemplo sencillo de dos variables que recoge las medidas de rigidez (X_1) y fuerza flexora (X_2), en unidades de $lb/inch^2$, de un tipo particular de sierra. Las especificaciones del proceso son:

$$\mu = \begin{pmatrix} 265 \\ 470 \end{pmatrix}, \ \mathbf{\Sigma} = \begin{pmatrix} 100 & 66 \\ 66 & 121 \end{pmatrix}$$

Para detectar las observaciones que están fuera de control aplicamos el gráfico de control T^2 con $\alpha = 0.05$. A continuación se trata de determinar si la causa de la situación fuera de control ha sido la variable X_1, la X_2 o ambas. Esto se puede ver como un problema de clasificación con tres clases: 'cambio en X_1', 'cambio en X_2' y 'cambio en ambas'. Considerando un cambio de 1.5σ, los vectores de medias son, respectivamente, $(265 + 1.5 * 10, 470)$, $(265, 470 + 1.5 * 11)$ y $(265 + 1.5 * 10, 470 + 1.5 * 11)$. De esta forma generamos 500 repeticiones cogiendo de cada una de ellas valores de la T^2 y de las

variables que marcaban las situaciones descritas anteriormente. Estas 1500 observaciones constituyen nuestro conjunto de entrenamiento.

Para probar el árbol de clasificación entrenado con 1.5σ, generamos, siguiendo el mismo procedimiento, conjuntos de observaciones con cambios de 2σ, 2.5σ y 3σ hasta conseguir 500 observaciones de cada clase.

Ejemplo 2: El caso de tres variables.

Este ejemplo se centra en una empresa que fabrica detergentes en la que se controlan tres variables: color, porcentaje libre de aceite y porcentaje de acidez. El problema de clasificación equivalente tiene $2^3 - 1 = 7$ clases. A partir de los datos muestrales tomados cada hora de producción, durante los diez primeros días, se estiman los valores de las medias y las varianzas y covarianzas, obteniendo los siguientes resultados:

$$\bar{\mathbf{x}} = \begin{pmatrix} 67.5 \\ 12.0 \\ 97.0 \end{pmatrix}, \ \mathbf{S} = \begin{pmatrix} 0.68 & 0.36 & -0.07 \\ & 1.00 & -0.12 \\ & & 0.03 \end{pmatrix}$$

Además, los datos siguen una distribución normal multivariante con las anteriores estimaciones para los parámetros. Para la condición de fuera de control mediante la T^2 de Hotelling se utiliza $\alpha = 0.05$, siendo el tamaño del cambio de 3σ para el conjunto de entrenamiento (700 observaciones) y de 2σ, 3σ y 4σ para el conjunto de test.

Ejemplo 3: El caso de cuatro variables.

Este ejemplo, previamente utilizado en Doganaksoy et al. (1991), utiliza cuatro variables relacionadas con pruebas de misiles. El vector de medias y la matriz de covarianzas son los siguientes:

$$\mu = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \ \mathbf{\Sigma} = \begin{pmatrix} 102.74 & 88.34 & 67.03 & 54.06 \\ & 142.74 & 86.55 & 80.02 \\ & & 64.57 & 69.42 \\ & & & 0 \end{pmatrix}$$

En este caso también se detecta la condición de fuera de control con la T^2 de Hotelling con $\alpha = 0.05$. Al trabajar con 4 variables aumenta considerablemente la complejidad del problema ya que son $2^4 - 1 = 15$ el número de posibles clases, siendo el tamaño del cambio de 2σ para el conjunto de entrenamiento (7500 observaciones) y de 2σ, 2.5σ y 3σ para el conjunto de test.

5 Experiencias computacionales: comparación de árboles de clasificación y redes neuronales

En el primer ejemplo, al utilizar dos variables para controlar el proceso, son tres las posibles clases a tener en cuenta. Una vez entrenado el árbol de clasificación sobre el conjunto generado con un cambio de 1.5 desviaciones típicas, se prueba su capacidad de generalización ante nuevas observaciones, en este caso con cambios de 2, 2.5 y 3 desviaciones típicas

respectivamente. Estos resultados pueden verse en la tabla 2. En los tres casos el árbol consigue claramente predecir la clase correcta en la mayoría de las ocasiones. Además, se aprecia como, a medida que aumenta el tamaño del cambio, es más sencillo para el clasificador detectar la verdadera causa de la señal fuera de control y comete un menor número de errores, salvo cuando el cambio se produce exclusivamente en la variable X_2 donde se producen dos fallos más cuando el cambio es de 3 desviaciones típicas que cuando éste es de 2.5 desviaciones, aunque en ambos casos es menor el número de errores que para un cambio de 2 desviaciones.

En el segundo ejemplo, el vector de medias y la matriz de covarianzas que utilizamos para generar los datos son estimaciones de los verdaderos valores de los parámetros que son desconocidos al tratarse de un caso real. Precisamente, la cercanía de este caso con la realidad lo hace aún más interesante. Al igual que ocurría en el ejemplo anterior, el árbol de clasificación es capaz de encontrar la verdadera naturaleza del cambio producido en la mayoría de los casos para los tres tamaños de cambio (2, 3 y 4 desviaciones típicas), excepto en el caso en el que el desplazamiento es de dos desviaciones típicas. En esta ocasión parece que el clasificador tiene dificultades para acertar cuando son varias las variables que han provocado la señal fuera de control (tabla 3).

El ejemplo 3 es, desde el punto de vista de los problemas de clasificación, el caso más complicado al tratarse de un número muy elevado de clases (15). A pesar de esta dificultad también el árbol consigue acertar en una amplia mayoría de los casos, tanto cuando el cambio se produce de forma aislada en una de las variables, como cuando son varias de ellas las que sufren una alteración. En general, el error es menor cuanto mayor es el cambio provocado en el vector de medias. La tabla 4 muestra los resultados para este ejemplo.

El objetivo de este trabajo no es establecer comparaciones con otros métodos ya utilizados sino presentar los árboles de clasificación como una herramienta útil para la interpretación de las señales fuera de control. Sin embargo, ya que los ejemplos utilizados para generar los datos de la aplicación han sido utilizados previamente por otros autores, nos parece conveniente presentar las comparaciones con los resultados de Niaki & Abbasi (2005). Esta comparación se realiza únicamente considerando los errores en los conjuntos de test en aras a una mayor brevedad; además, deben hacerse con la necesaria cautela ya que no son exactamente los mismos datos sino que se han generado utilizando los mismos vectores de medias y matrices de covarianzas. La tabla 1 recoge los porcentajes de error del método que proponemos en este trabajo, árbol de clasificación (CART), y de los métodos utilizados por los autores citados anteriormente, perceptron multicapa (MLP) y gráficos de control Shewhart multivariante (MSCH). Puede observarse como el árbol de clasificación supera en todos los casos al MLP y en todos menos uno al MSCH. Es decir, sólo en el ejemplo de dos variables y cuando el cambio provocado es de dos desviaciones típicas no es el árbol de clasificación el método ganador.

Método	**2 variables**			**3 variables**			**4 variables**		
	2σ	2.5σ	3σ	2σ	3σ	4σ	2σ	2.5σ	3σ
CART	9.07	6.53	5.47	49.86	12.71	18.71	15.28	14.08	11.19
MLP*	13.93	10.73	8.73	50.00	62.14	45.29	33.03	24.87	18.69
MSCH*	4.60	10.93	10.93	n.d.	n.d.	n.d.	71.15	66.68	68.53

Tabla 1: Porcentaje de error en los conjuntos de test (*Niaki & Abbasi, 2005).

6 Conclusiones

En este trabajo hemos propuesto el uso de los árboles de clasificación para la interpretación de las señales fuera de control ocurridas en el control de la calidad de procesos multivariantes. Los árboles de clasificación son una herramienta muy potente, flexible y sencilla de interpretar ya que se puede seguir el proceso que sigue una observación hasta el nodo hoja donde es clasificada. La aplicación práctica realizada pone de manifiesto la utilidad de los árboles para esta tarea con resultados muy esperanzadores en todos los ejemplos y bajo todos los supuestos contemplados. Además, el uso diferenciado de muestras de entrenamiento y test garantiza la capacidad de generalización de estos clasificadores. Esto significa que el clasificador será capaz de clasificar de forma correcta nuevas observaciones ante distintas situaciones.

Son muchas las cuestiones que no hemos abarcado en este trabajo pero que nos gustaría llevar a cabo en trabajos futuros, como por ejemplo, utilizar los árboles de clasificación no sólo como herramienta para interpretar las señales fuera de control sino también para intentar detectar si el proceso se encuentra o no bajo control o incluso el tamaño del cambio producido.

Referencias

[1] Alt, F.B. (1985) "Multivariate quality control" , en: N.L. Johnson & S. Kotz (Eds.) *Encyclopedia of Statistical Sciences* vol. **6**, John Wiley & Sons, New York.

[2] Aparisi, F.; Avendaño, G.; Sanz, J. (2006) "Techniques to interpret T^2 control chart signals", *IIE Transactions* **38**: 647–657.

[3] Breiman, L.; Friedman, J.H.; Olshen, R.; Stone, C.J. (1984) *Classification and Regression Trees*. Wadsworth International Group, Belmont.

[4] Chang S.I.; Aw C.A. (1996) "A neural fuzzy control chart for detecting and classifying process mean shifts", *International Journal of Production Research* **34**(8): 2265–2278.

[5] Cheng, C.S. (1995) "A multi-layer neural network model for detecting changes in the process mean", *Computers and Industrial Engineering* **28**(1): 51–61.

[6] Cheng, C.S. (1997) "A neural network approach for the analysis of control chart patterns", *International Journal of Production Research* **35**(3): 667–697.

[7] Doganaksoy, N.; Faltin, F.W.; Tucker W.T. (1991) "Identification of out of control quality characteristics in a multivariate manufacturing enviroment", *Communications in Statistics - Theory and Methods* **20**: 2775–2790.

[8] Guh, R.S. (2003) "Integrating artificial intelligence into on-line statistical process control" , *Quality and Reliability Engineering International* **19**: 1–20.

[9] Guh, R.S. (2007) "On-line identification and quantification of mean shifts in bivariate processes using a neural network-based approach", *Quality and Reliability Engineering International* **23**: 367–385.

[10] Guh, R.S.; Tannock, J.D.T. (1999) "A neural network approach to characterize pattern parameters in process control charts", *Journal of Intelligent Manufacturing* **10**(5): 449–462.

[11] Hayter, A.J.; Tsui, K.L. (1994) "Identification and quantification in multivariate quality control problems", *Journal of Quality Technology* **26**: 197–208.

[12] Jackson, J.E. (1980) "Principal components and factor analysis: Part I - Principal components", *Journal of Quality Technology* **12**: 201–213.

[13] Mason, R.L.; Young, J.C. (2002) "Multivariate statistical process control with industrial applications" , *American Statistical Association and the Society for Industrial and Applied Mathematics (ASA-SIAM)*: Philadelphia.

[14] Mason, R.L.; Tracy, N.D.; Young, J.C. (1995) "Decomposition of T^2 for multivariate control chart interpretation", *Journal of Quality Technology* **27**: 109–119.

[15] Murphy, B.J. (1987) "Selecting out of control variables with the T^2 multivariate quality control procedure" , *Journal of the Royal Statistical Society* **36** Serie D (The Statistician): 571–581.

[16] Niaki, S.T.A.; Abassi, B. (2005) "Fault diagnosis in multivariate control charts using artificial neural networks", *Quality and Reliability Engineering International* **21**: 825–840.

[17] Noorossana, R.; Farrokhi, M.; Saghaei, A. (2003) "Using neural networks to detect and classify out-of-control signals in autocorrelated processes", *Quality and Reliability Engineering International* **19**: 1–12.

[18] Quinlan, J.R. (1986) "Induction on decision trees", *Machine Learning* **1**: 81–106.

[19] Quinlan, J.R. (1993) *C4.5 Programs for Machine Learning*. Morgan Kaufmann, San Mateo, CA.

[20] Tracy, N.D.; Young, J.C.; Mason, R.L. (1992) "Multivariate control charts for individual observations", *Journal of Quality Technology* **24**: 88–95.

[21] Zorriassatine, F.; Tannock, J.D.T. (1998) "A review of neural networks for statistical process control", *Journal of Intelligent Manufacturing* **9**(3): 209–224.

Apéndice

Las tablas siguientes recogen las matrices de confusión (tres en cada tabla) para cada uno de los casos considerados. Por ejemplo, para el caso de dos variables con un cambio de 2 desviaciones típicas (2σ)de las 500 observaciones generadas con cambio en las dos variables (X_1, X_2), 425 han sido asignadas correctamente por el árbol y 75 no, puesto que ha considerado un cambio en X_1 en 52 casos y en X_2 para los 23 restantes. De forma similar se pueden interpretar el resto de filas para cada matriz de confusión. En cada una de éstas, los valores de la diagonal principal recogen los casos asignados de forma correcta.

		Clase asignada								
		cambio de 2σ			cambio de 2.5σ			cambio de 3σ		
	Variables	X_1	X_2	X_1, X_2	X_1	X_2	X_1, X_2	X_1	X_2	X_1, X_2
Clase obs	X_1	478	0	22	482	0	18	492	0	8
	X_2	0	461	39	0	466	34	0	464	36
	X_1, X_2	52	23	425	32	14	454	23	15	462

Tabla 2: Matrices de confusión para distintos cambios en 2 variables.

		Clase asignada (cambio de 2σ)						
		X_1	X_2	X_3	X_1, X_2	X_1, X_3	X_2, X_3	X_1, X_2, X_3
Clase observada	X_1	80	0	12	8	0	0	0
	X_2	0	71	26	3	0	0	0
	X_3	6	4	83	6	1	0	0
	X_1, X_2	14	16	16	54	0	0	0
	X_1, X_3	38	1	29	6	26	0	0
	X_2, X_3	0	47	32	1	0	18	2
	X_1, X_2, X_3	2	11	13	31	19	5	19
		Clase asignada (cambio de 3σ)						
		X_1	X_2	X_3	X_1, X_2	X_1, X_3	X_2, X_3	X_1, X_2, X_3
Clase observada	X_1	94	0	1	4	1	0	0
	X_2	1	87	5	7	0	0	0
	X_3	3	3	87	3	3	0	1
	X_1, X_2	7	3	3	84	3	0	0
	X_1, X_3	5	0	10	0	80	0	5
	X_2, X_3	0	5	1	0	0	87	7
	X_1, X_2, X_3	0	0	0	2	1	5	92
		Clase asignada (cambio de 4σ)						
		X_1	X_2	X_3	X_1, X_2	X_1, X_3	X_2, X_3	X_1, X_2, X_3
Clase observada	X_1	88	0	1	5	6	0	0
	X_2	0	64	0	4	0	31	1
	X_3	1	4	81	0	3	10	1
	X_1, X_2	0	1	0	72	0	1	26
	X_1, X_3	0	0	1	0	75	0	24
	X_2, X_3	0	0	0	0	0	89	11
	X_1, X_2, X_3	0	0	0	0	0	0	100

Tabla 3: Matrices de confusión para distintos cambios en 3 variables.

Clase asignada (cambio de 2σ)

Clase observada	X_1	X_2	X_3	X_4	X_1,X_2	X_1,X_3	X_1,X_4	X_2,X_3	X_2,X_4	X_3,X_4	X_1,X_2,X_3	X_1,X_2,X_4	X_1,X_3,X_4	X_2,X_3,X_4	Todas
X_1	417	0	0	1	12	17	29	0	0	0	3	2	9	0	10
X_2	0	414	0	0	22	0	0	11	29	0	6	4	0	8	6
X_3	0	0	404	0	0	30	0	17	0	35	10	0	0	3	1
X_4	0	0	0	442	0	0	19	0	14	8	0	3	5	5	4
X_1,X_2	8	11	0	0	433	0	0	0	2	0	6	32	0	0	8
X_1,X_3	17	0	16	0	0	429	0	1	0	1	11	0	24	0	1
X_1,X_4	19	0	0	26	0	0	419	0	0	0	0	15	15	0	6
X_2,X_3	0	9	9	0	0	0	0	429	0	0	30	0	0	23	0
X_2,X_4	0	14	0	20	2	0	1	0	440	0	0	13	0	14	6
X_3,X_4	0	0	19	8	0	1	0	0	0	444	0	0	17	10	1
X_1,X_2,X_3	3	6	9	0	11	18	0	14	0	0	423	0	1	3	12
X_1,X_2,X_4	7	2	0	11	26	0	17	0	41	0	0	390	0	0	6
X_1,X_3,X_4	13	0	2	7	0	12	5	0	0	14	0	0	437	0	10
X_2,X_3,X_4	0	8	6	4	0	0	0	15	8	8	1	0	0	443	7
Todas	9	1	0	7	14	3	7	1	7	7	11	15	11	17	390

Clase asignada (cambio de 2.5σ)

Clase observada	X_1	X_2	X_3	X_4	X_1,X_2	X_1,X_3	X_1,X_4	X_2,X_3	X_2,X_4	X_3,X_4	X_1,X_2,X_3	X_1,X_2,X_4	X_1,X_3,X_4	X_2,X_3,X_4	Todas
X_1	397	0	0	1	19	18	43	0	0	0	4	4	12	0	2
X_2	0	402	0	0	38	0	0	11	25	0	4	5	0	13	2
X_3	0	0	393	0	0	41	0	16	0	46	2	0	1	1	0
X_4	0	0	0	404	0	0	37	0	22	12	0	8	11	3	3
X_1,X_2	3	8	0	0	443	0	0	0	0	0	5	38	0	0	3
X_1,X_3	7	0	9	0	0	458	0	0	0	0	15	0	11	0	0
X_1,X_4	13	0	0	13	0	0	450	0	0	0	0	10	12	0	2
X_2,X_3	0	6	10	0	0	0	0	441	0	0	17	0	0	26	0
X_2,X_4	0	7	0	8	4	0	1	0	446	0	0	24	0	9	4
X_3,X_4	0	0	8	4	0	1	0	0	0	447	0	0	22	18	0
X_1,X_2,X_3	5	1	9	0	5	10	0	43	0	0	415	0	0	2	10
X_1,X_2,X_4	1	0	0	4	10	0	13	0	18	0	0	452	0	1	1
X_1,X_3,X_4	3	0	1	4	0	22	4	0	0	15	0	0	445	0	6
X_2,X_3,X_4	0	3	2	3	0	0	0	13	6	10	0	0	0	454	9
Todas	4	0	1	4	8	4	5	0	4	10	13	9	12	29	397

Clase asignada (cambio de 3σ)

Clase observada	X_1	X_2	X_3	X_4	X_1,X_2	X_1,X_3	X_1,X_4	X_2,X_3	X_2,X_4	X_3,X_4	X_1,X_2,X_3	X_1,X_2,X_4	X_1,X_3,X_4	X_2,X_3,X_4	Todas
X_1	412	0	0	0	13	23	36	0	0	0	5	3	7	0	1
X_2	0	410	0	0	31	0	0	8	29	0	4	3	0	13	2
X_3	0	0	397	0	0	36	0	23	0	36	2	0	2	4	0
X_4	0	0	0	395	0	0	49	0	22	14	0	10	5	2	3
X_1,X_2	0	3	0	0	452	0	0	1	0	1	4	37	0	0	3
X_1,X_3	10	0	1	0	0	468	0	1	0	1	11	0	7	0	1
X_1,X_4	5	0	0	4	0	0	455	0	0	0	0	11	23	0	2
X_2,X_3	0	1	3	0	0	0	0	468	0	0	6	0	0	22	0
X_2,X_4	0	5	0	1	0	0	0	0	463	0	0	19	0	11	1
X_3,X_4	0	0	1	1	0	0	0	0	0	475	0	0	15	8	0
X_1,X_2,X_3	0	1	6	0	0	6	0	70	0	0	413	0	0	0	4
X_1,X_2,X_4	0	0	0	0	7	0	4	0	10	0	0	474	0	0	5
X_1,X_3,X_4	2	0	0	1	0	9	4	0	0	8	0	0	472	0	4
X_2,X_3,X_4	0	3	4	0	0	0	0	7	4	2	0	0	0	475	5
Todas	3	0	0	4	4	1	5	0	1	2	12	8	11	17	432

Tabla 4: Matrices de confusión para distintos cambios en 4 variables.