Caracterización de atributos en Data Mining a través del método Ks-Monte Carlo-Bootstrap

Jorge I. Paolini R.

Departamento de Ciencia y Tecnología
Universidad Nacional Experimental de Guayana
Puerto Ordaz, Venezuela
Grupo de Investigación en Simulación, Optimización y Muestreo
e-mail: jpaolini@cantv.net

RESUMEN

Se presenta el Método KS-Monte Carlo-Bootstrap para la caracterización de atributos en Data Mining. A partir de la ejecución del algoritmo Bootstrap y la inserción de la prueba Kolmogorov- Smirnov (KS) en el algoritmo se obtiene la cantidad de probabilidad G de una función de densidad $f(x;\theta)$ que posee el espacio muestral de las muestras generado. Se ejemplifica la aplicación del método con la caracterización de dos atributos relevantes en la operación de celdas electrolíticas P19 para la producción de aluminio primario.

Palabras claves: Bootstrap, Monte Carlo, Kolmogorov-Smirnov, Data Mining.

ABSTRACT

CHARACTERIZATION OF ATTRIBUTES IN DATA MINING THROUGH METODO KS-MONTE- CARLO-BOOTSRAP

A Bootstrap Monte Carlo Method is presented for variable identification in Data Mining. Starting with the application of Bootstrap and the insertion of the Kolmogorov-Smirnov test (KS) in the algorithm, a probability G is obtained of a density function $f(x;\theta)$ which has the resampling space generated by the Bootstrap method. There are examples of the application of the proposed method through the characterization of two relevant variables within the operation of P19 electrolytic pot cells of primary aluminum production.

Keywords: Bootstrap, Monte Carlo, Kolmogorov-Smirnov, Data Mining.

Artículo recibido el 23 de Noviembre de 2006 y aceptado en su forma final el 13 de Diciembre de 2006

I. INTRODUCCIÓN

El método Bootstrap permite la construcción del espacio muestral de las muestras, independientemente de la asunción de normalidad o de una densidad $f(x;\theta)$ conocida. Asumir la normalidad puede conducirnos a equívocos en la determinación de estimaciones acerca de un atributo X, de tal modo que prescindir de este supuesto conduce a elaborar un cuerpo de métodos que sean "libres" de alguna distribución en el atributo que se esté observando. Se presenta un método para obtener la cantidad de probabilidad que posee un atributo X respecto de una distribución de probabilidades $f(x;\theta)$. Estamos interesados en determinar cuanta cantidad de probabilidad de una densidad $f(x;\theta)$ conocida posee un atributo. Se conoce el espacio muestral de las muestras generado por simulación de Monte Carlo siguiendo el esquema del Método Bootstrap. Conocemos los valores de un atributo $X = \{X1, X_2,..., Xn\}$ ¿Cómo encontramos la proporción de una distribución $F(x;\theta)$ para el conjunto de valores del atributo X ?

Este problema no se ha encarado y usualmente suponemos normalidad en el atributo considerado o simplemente no hacemos suposición alguna cuando realizamos estimaciones y otras inferencias acerca de la naturaleza del mismo.

II. DESARROLLO

1. Un Método para Caracterizar Atributos a partir del Método Ks-Monte Carlo-Bootstrap

En el Método Bootstrap subyace la idea de generar un espacio muestral de las muestras, el algoritmo para la caracterización de un atributo X se muestra a continuación:

- Determinar F_n la función de distribución de los n datos observados, asignando probabilidad 1/n a cada valor del atributo $X = \{ X_1, X_2, ... , X_n \}$.
- Con un generador de números aleatorios uniformes e independientes en [1,n] tomar n nuevos datos con reemplazo de F_n, para obtener la muestra del atributo: $X^* = \{ X_1^*, X_2^*, ... , X_n^* \}$, a esta muestra se le denomina muestra Monte Carlo-Bootstrap.
- Aplicar la prueba *KS* en la muestra Monte Carlo-Bootstrap X^* para determinar el valor de aceptación o rechazo y^*.
- Repetir los pasos (2) y (3) un gran número de veces, digamos B veces.

La serie de valores generada durante la ejecución del algoritmo será $y^*_1, y^*_2, ..., y^*_b, ..., y^*_B$, con estos B valores se determina cuanta cantidad de una distribución $F(x;\theta)$ posee el atributo X. En cada iteración del algoritmo se obtiene un valor del estadístico y^*_b para cada conjunto de valores del atributo X^*_b, así la aceptación o rechazo de cada conjunto del espacio muestral de las muestras se obtiene a medida que el algoritmo se realiza**.**

Para caracterizar el espacio muestral generado en cada ciclo que se ejecuta el algoritmo Bootstrap se aplica la prueba *KS*, entonces, se realizan B pruebas *KS* a cada conjunto del espacio muestral de las muestras. Se obtiene en cada realización de un ciclo *KS-Bootstrap* y un resultado de la prueba: rechazo o aceptación, cada resultado lo denotaremos como un valor de una variable aleatoria Bernoulli Y, conviene que cuando la prueba *KS* se rechace el valor que toma la variable Y sea 0 y cuando la prueba *KS* se acepte el valor será $Y=1$. Entonces tendremos la sucesión de valores $y_1^*, y_2^*, ... , y_b^*, ... , y_B$ de Y que determinará la cantidad de probabilidad que posee el espacio muestral de las muestras de la función de distribución $F(x;\theta)$ que se fijó en la *Ho* de la prueba *KS*.

La cantidad de probabilidad de $f(x;\theta)$ contenida en el atributo X será $G_X=\Sigma y_i^*/B$, esta cantidad nos indica la proporción de F contenida en el atributo considerado y nos determina la caracterización de X bajo la Hipótesis *Ho* en la prueba *KS* de cada ejecución del algoritmo Bootstrap.

2. Caracterización de Atributos en la Operación de las Celdas de Producción de Aluminio P19

En la producción de aluminio primario existen una gran cantidad de atributos asociados con la operación del proceso. Uno de los factores que hemos encontrado que alteran el proceso es la humedad relativa en el ambiente debida a los períodos de lluvia. Se han establecido dos períodos de análisis dependiendo de la humedad: seco y lluvioso. Para este trabajo hemos tomado la Temperatura del proceso y la Acidez (exceso de AlF_3).

Pueden observarse en el gráfico 1 las densidades para la temperatura. La cantidad de normalidad o gaussianidad *G* ($0 \leq G \leq 1$) del atributo Temperatura en el proceso electrolítico se muestra en la tabla I. La cantidad *G* fue obtenida con *B*=5000 iteraciones del algoritmo y con el promedio de diez replicas del experimento.

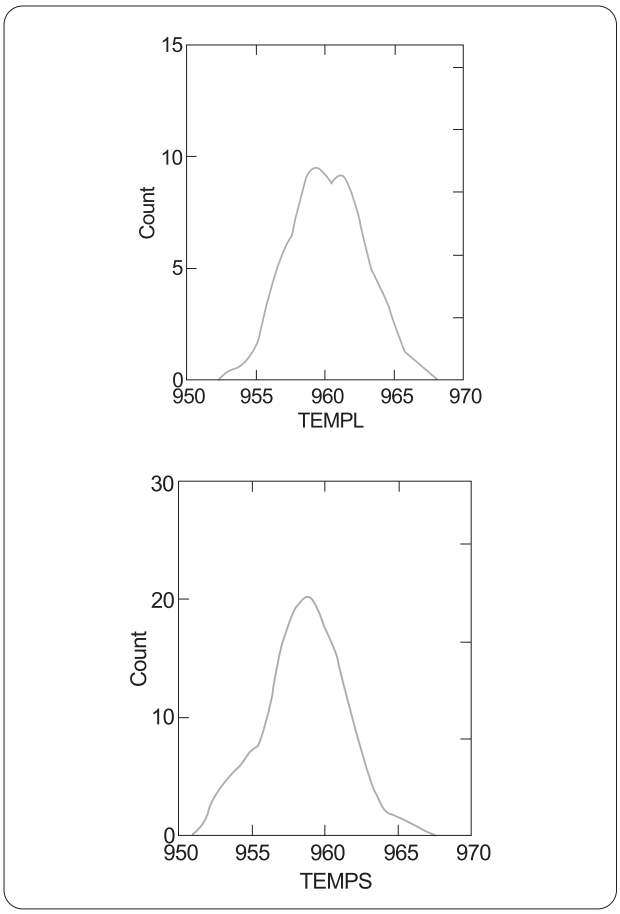

Gráfico 1. Densidad Kernel de la Temperatura en los períodos lluvioso (L) y seco (S)

Nótese que a pesar de que los coeficientes de variación del atributo son muy parecidos, la cantidad de normalidad G en cada período difiere $G_{TempL} < G_{TempS}$. Utilizando la prueba de proporciones se determina que hay diferencias significativas entre los valores de la caracterización $G_{Temp|Normal}$, es decir, la cantidad de Normalidad que posee el atributo Temperatura en cada período difiere significativamente (p=0,00003).

Tabla I. Caracterización de la Temperatura en los períodos lluvioso y seco

	Período Lluvioso (L) n=28	Período Seco n=82
Normalidad *G*	*0,68442*	*0,97030*
Promedio μ	*960,13*	*958,51*
LI μ(1-α = 0,95)	*959,18*	*957,93*
LS μ(1-α = 0,95)	*961,08*	*959,09*
Asimetría	*0,0073*	*-0,0110*
Coeficiente de variación	*0,0027*	*0,0028*

En el grafico 2 se pueden observar las densidades para el atributo acidez, es de observar la notable asimetria de la densidad durante el periodo lluvioso (acidez L).

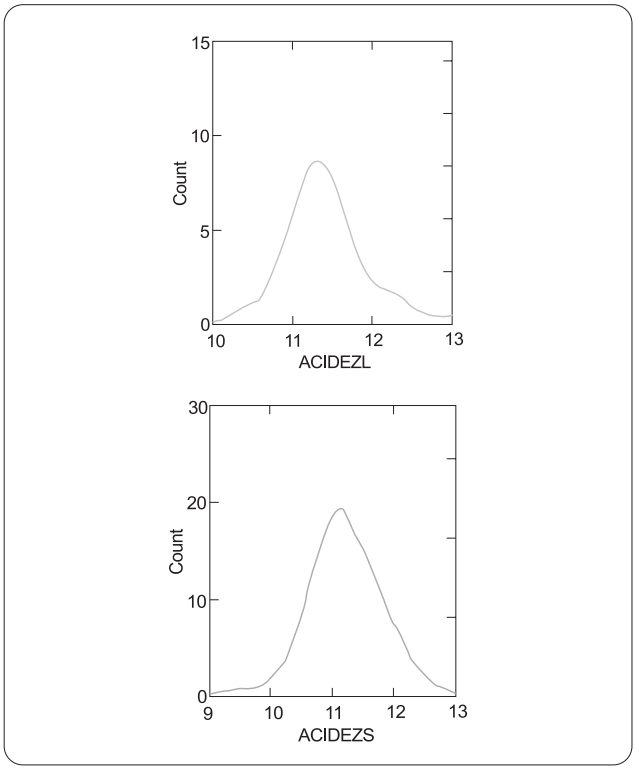

Gráfico 2. Densidad Kernel de la Acidez en los períodos lluvioso (L) y seco (S)

El exceso de Fluoruro de Aluminio se mide con el atributo denominado Acidez, el consumo de este electrolito aumenta en el período lluvioso, en la tabla II podemos ver un leve aumento de la acidez durante este período.

Tabla II. Caracterización de la Acidez en los períodos lluvioso y seco

	Período Lluvioso (L) n = 28	Período Lluvioso (S) n = 82
Normalidad *G*	0,29374	0,87348
Promedio μ	11,4321	11,2128
LI $\mu(1-\alpha = 0,95)$	11,2571	11,0870
LS $\mu(1-\alpha = 0,95)$	11,6214	11,3387
Asimetría	0,6647	0,2209
Coeficiente de variación	0,0443	0,0518

Cuando caracterizamos la Acidez del proceso encontramos una falta de gaussianidad durante el período lluvioso ($1-G_{AcidezL}=0,70626$). Si calculamos estimaciones para este atributo haciendo la suposición de Normalidad podemos cometer el error de sobreestimarlo o subestimarlo, como puede comprobarse en la tabla III.

Tabla III. Estimaciones para la Acidez en el período lluvioso

	Método Normal (n = 28)	Método Bootstrap (B = 5000)
Promedio μ	11,4321	11,4393
LI $\mu(1-\alpha = 0,95)$	11,2446	11,2571
LS $\mu(1-\alpha = 0,95)$	11,6197	11,6214
L (Incertidumbre)	0,37510	0,36430

En la tabla podemos observar que debido a la suposición de normalidad se subestima el valor de la Acidez promedio en el proceso electrolítico, la consecuencia natural de esta subestimación es a su vez la subestimación de la cantidad de AlF_3 que se debe adicionar para que el proceso de producción de aluminio primario sea estable durante el período de lluvias. Nótese que la incertidumbre en la estimación por el Método Normal es mayor que en el Método Bootstrap (este método no supone para el cálculo de las estimaciones alguna distribución para el atributo Acidez).

III. CONCLUSIONES

Dado que se asume normalidad en muchos atributos que provienen de procesos de medición, la caracterización bajo la suposición de gaussianidad es obvia. Sin embargo, para mediciones de tiempos y movimientos la suposición de exponencialidad es recomendable.

Cuando la suposición de normalidad en la estimación de un atributo que posee una $G_{X|Normal}$ baja ($G < 0,5$) nos puede conducir a equívocos, entonces se debe prescindir de esta suposición y optar por procedimientos de estimación que presupongan poco o nada acerca de la distribución del atributo en cuestión.

La caracterización de atributos en Data Mining permite determinar la cantidad de probabilidad G que posee un atributo X dada una distribución $f(x;\Theta)$. Este método permite obtener información útil del espacio muestral de las muestras para la toma de decisiones en ambientes industriales.

IV. BIBLIOGRAFÍA

- Conover, W.J., *Practical Nonparametric Statistics*. Wiley Series in Probability and Statistics.Second Edition. New York, (1999).
- Chernick, M., *Bootstrap Methods, a Practitioner's Guide*. Wiley Interscience Publication. New York, (1999).
- Chskin, D., *Parametric and Nonparametric Statistical procedures*.Chapman & Hall/CRC. Second Edition. Boca Raton, (2000).
- Efron, B and Tibshirani, R., *An Introduction to the Bootstrap*. Chapman & Hall/CRC. Boca Raton, (1998).
- Paolini, J., *El Método Bootstrap: un paradigma en la formación de los ingenieros en Computación, Informática y Sistemas*. Actas de la III Conferencia Latinoamericana de Facultades de Ingeniería y Escuelas de Ingeniería de Sistemas y Ciencias de la Computación CONLATI 99. Barquisimeto, Venezuela, (1999).
- Paolini, J. *Fundamentos del Método Bootstrap*. Corporación Aluminios de Venezuela. Tutorial en la Conferencia Latinoamericana de Facultades de Ingeniería y Escuelas de Ingeniería de Sistemas y Ciencias de la Computación. Barquisimeto, Venezuela, (1999).
- Paolini, J., *Caracterización de Poblaciones Estadísticas a través del Método Bootstrap*. Ponencia invitada de la III Asamblea del Departamento de Ciencia y Tecnología. Universidad Nacional Experimental de Guayana. Trabajo no publicado. Venezuela, (2000).
- Paolini, J., *Métodos de Estimación y Estimación Aproximada*. Disertación para optar al cargo de Profesor Agregado. Mimeografiado Universidad Nacional Experimental de Guayana. Venezuela. (2001).
- Paolini, J., El Método de Monte Carlo-Bootstrap como un modelo de enseñanza en la solución del problema de Estimación. En Memorias de RELME 19. Uruguay, (2005).

Sitios Web

- Halbert White. *Reality Check for Data Mining.* [En línea]. Disponible: http://secondmoment.org/articles/datamining.php. 05 Agosto 2005.
- Jagoda Crawford & Frank Crawford. *Data Mining in a Scientific Environment.* [En línea] Disponible: http:// csu.edu.au/special/auugwww96/proceedings/crawford/crawford.html. 10 de Agosto 2005.
- John Maindonald. *Data Mining from a Statistical Perspective.* [En línea]. Disponible: http:// wwwmaths.anu.edu.au/~johnm/dm/dmpaper.html. 05 de Agosto 2005.

POSTCRIPTUM

Para la selección de los dos atributos en este trabajo se utilizó un modelo Logit (véase el anexo). Los dos mejores descriptores de los períodos lluvioso y seco son la Temperatura del proceso y la Acidez del baño electrolítico. El modelo Logit que se ajusta a esta situación se expresa según la siguiente ecuación

$$\gamma_{LLUVIA} = -601{,}014 + 0{,}592 \cdot \mathbf{Temp} + 2{,}835 \cdot \mathbf{Acidez}$$

El Logaritmo de la verosimilitud (Likelihood Log) es 2*[LL(n)-LL(0)] = 26.585, para dos grados de Libertad el valor de Ji-Cuadrado observado se corresponde con un valor de p = 0.000.

ANEXO

REGRESION LOGIT

SYSTAT Rectangular File C:\Mis Documentos**DAT_AEMP_L3-Acidez_Temp.SYD**,
Created Mon Aug 15, 2005 at 11:52:25, contains variables:

TEMPL NIVML ACIDEZL SOLUBL TEMPS NIVMS
ACIDEZS NA2OS HUMS ESTAC$ **TEMP** **ACIDEZ**
LLUVIA

Categorical values encountered during processing are:
LLUVIA (2 levels) **0, 1**
Categorical variables are effects coded with the highest value as reference.

Binary LOGIT Analysis.

Dependent variable: LLUVIA
Input records: 82
Records for analysis: 82
Sample split

Category choices
REF **28 (LLUVIA = 1)**
RESP **54 (LLUVIA = 0)**
Total : **82**

L-L at iteration 1 is -56.838
L-L at iteration 2 is -40.832
L-L at iteration 3 is -39.420
L-L at iteration 4 is -39.352
L-L at iteration 5 is -39.352
L-L at iteration 6 is -39.352
Log Likelihood: -39.352

Parameter	Estimate	S.E.	t-ratio	p-value
1 CONSTANT	-601.014	147.511	-4.074	0.000
2 TEMP	0.592	0.147	4.028	0.000
3 ACIDEZ	2.835	0.777	3.647	0.000

		95.0 % bounds	
Parameter	Odds Ratio	Upper	Lower
2 TEMP	1.808	2.412	1.355
3 ACIDEZ	17.024	78.114	3.710

Log Likelihood of constants only model = LL(0) = -52.644
2*[LL(N)-LL(0)] = 26.585 with 2 df Chi-sq **p-value = 0.000**