Caracterización de Tráfico-Distribución de Johnson SB

Luis Marrone

Calle 50 y 120, 2do piso, La Plata, Buenos Aires, Argentina
Laboratorio de Investigación en Nuevas Tecnologías Informáticas
Facultad de Informática-Universidad Nacional de La Plata
lmarrone@linti.unlp.edu.ar

Resumen. Obtener el modelo de tráfico resulta crucial a la hora de evaluar la performance de una red como así también disponer del mismo en la etapa de desarrollo e implementación de la misma. El punto de partida para obtener el modelo es contar con la caracterización del tráfico a cursar por la red. Caracterizar el tráfico redunda en obtener la distribución que mejor se corresponda con el mismo ("fitting distribution"). En este trabajo presentamos una distribución, "Johnson SB" normalmente no utilizada en la construcción de estos modelos pero que como veremos presenta resultados promisorios.

Palabras clave: Redes. Tráfico. Modelos. Distrtibuciones.Validación

1. Introducción

Caracterizar el tráfico de una red es proveer la descripción completa de los elementos que lo definen como ser entre otros, el tiempo de arribo entre mensajes, su longitud y el tiempo de servicio. Para disponer de esos elementos como paso previo a su descripción es que tomamos muestras. Muestras que se corresponden a variables aleatorias por lo cual no nos queda otro remedio que hacer un análisis estadístico de las mismas. La estadística descriptiva es la disciplina que nos otorga ese análisis en una primera etapa, dándonos valores del promedio, varianza, desviación estándar, curtosis y asimetría. Estas propiedades de las muestras tomadas nos ayudan a elegir las distribuciones más adecuadas que se correspondan con las muestras. El promedio y la desviación estándar junto con la varianza están relacionados en cuanto al aporte dado que valores de la desviación estándar cercanos al promedio poco nos pueden aportar para la elección de la distribución, dada la dispersión que presentan las muestras. Tal vez un mayor aporte lo representan la asimetría y kurtosis aunque normalmente son poco empleadas en este tipo de análisis. En esta primera etapa también se suele acudir a representaciones gráficas de las variables, en particular el histograma que resulta valioso por cuanto permite acotar rápidamente un conjunto posible de distribuciones adecuadas. Obteniendo así como resultado de la primera etapa un conjunto de distribuciones posibles pasamos a la segunda etapa donde parametrizamos las distribuciones que elegimos en la primera y comparamos su comportamiento con el de las muestras tomadas. Esta etapa normalmente se conoce como adaptación/ajuste/sintonía de la distribución a las muestras ("fitting distribution"). Esta

sintonía se puede realizar en forma gráfica y/o analítica, de hecho se complementan. Los métodos gráficos tradicionales son el de Q-Q-Plot y P-P-Plot. Q-Q-Plot es un gráfico que compara dos distribuciones de probabilidades, la empírica obtenida a partir de las muestras y la teórica a partir de las distribuciones elegidas como las más adecuadas en la primer etapa y cuyos parámetros fueron estimados a partir de las muestras. Específicamente se grafican los percentiles en Q-Q plot donde la mayor correspondencia ente ambas distribuciones se dará si la gráfica resultantes es la recta y=x. Para el caso de P-P Plot se grafican ambas funciones de distribución acumulativas donde la mayor correspondencia se dará si la gráfica resultante se da con una pendiente d 45°. El método analítico comprende un conjunto de tests bajo el nombre de Bondad de Ajuste. Los tests comúnmente empleados son el de Kolmogorov-Smirnov, Anderson-Darling y Chi-Cuadrado.

En lo que sigue (Sección 2) detallamos elementos de la estadística descriptiva mencionados anteriormente junto con la distribución de Johnson SB [1, 2, 3], comúnmente no empleada en el análisis de tráfico. En la Sección 3 detallamos el escenario en el que se tomaron las muestras y la metodología empleada para caracterizarlo, continuando (Sección 4) con los resultados y conclusiones (Sección 5)

2. Elementos de Estadística

Detallamos aquí elementos de estadística descriptiva que colaboran eficazmente a la tarea de elegir posibles distribuciones que se correspondan con la realidad.

2.1. Curtosis[4]

La curtosis determina el grado de concentración/amplitud de pico que presentan las muestras en la región central de la distribución, comparada contra una normal. Así puede ser:

Leptocúrtica.- Existe una gran concentración. Supera a la normal, valores positivos

Mesocúrtica.- Existe una concentración normal. Valor cero

Platicúrtica.- Existe una baja concentración. Está por debajo de una normal. Valores negativos.

Para una muestra de n valores el coeficiente de curtosis está dado por:

$$g_2 = \frac{m_4}{m_2^2} - 3 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^2} - 3 \quad (1)$$

Los m_i corresponden a los momentos de orden i. A la relación de momentos se le resta 3 por cuanto es el valor que corresponde a una distribución normal.

2.2. Asimetría ("skewness")

Nos da una medida de la inclinación hacia derecha (>0) o izquierda (<0) de una distribución comparada con la normal.

Para una muestra de n valores la asimetría está dada por:

$$g_1 = \frac{m_3}{m_2^{3/2}} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^3}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2\right)^{3/2}} \quad (2)$$

2.3. Distribución de Johnson SB

En realidad constituyen un conjunto de cuatro distribuciones, "SL" o log-normal, "SU" sin límites, "SB", acotada y "SN", el caso particular de una normal. Creada por Norman Johnson en 1940 quien la creó con el objeto de aplicar los métodos y teoría de la distribución normal a un amplio rango de distribuciones no normales a través de transformaciones computables a partir de distribuciones como la exponencial y seno hiperbólico. Dada su flexibilidad esta familia de distribuciones se emplea en varios campos como el de química atmosférica [5], ingeniería biomédica [6], economía [7], gerenciamiento [8] , ciencia de los materiales [9] y análisis forestal[10], [11].

La distribución se define como:

$$f(x) = \frac{\delta}{\lambda\sqrt{2\pi}z(1-z)} e^{-\frac{1}{2}(\gamma+\delta\ln(\frac{z}{1-z}))^2}, \quad z \equiv \frac{x-\zeta}{\lambda} \quad (3)$$

Donde:
γ: Factor de forma
δ: Factor de forma (δ > 0)
λ: Factor de escala (γ > 0)
ζ : Factor de Localización
No tenemos registro de su empleo en el análisis de tráfico de redes de datos.

3. Escenario y metodología empleados

El escenario de prueba fue extremadamente simple por cuanto se quiso ver la factibilidad del empleo de la distribución de Johnson SB en redes de datos. Se realizó la captura con Wireshark 2.0.3 de sesiones de https entre una PC del laboratorio con el servidor del proyecto Gutenberg (http://www.gutenberg.org/ebooks/author/85)

De las muestras obtenidas se tomó el tiempo de servicio de cada segmento de datos recibido. Ese parámetro va a ser la componente de tráfico a caracterizar.

Aisladas entonces las muestras de ese tiempo se procedió a estimar los parámetros de distribuciones Normal, Beta y Johnson a los efectos de comparación con la nueva distribución propuesta.

Dados los estimadores se realizaron los tests de Kolmogorov y Anderson Darling para la Bondad de Ajuste.

Los cálculos estadísticos fueron realizados con Matlab R2017a y la Bondad de Ajuste con Johnson Curve Toolbox for Matlab [12]

4. Resultados

Presentamos los resultados producto del ajuste de las muestras a las distribuciones Normal, Beta y JohnsonSB

4.1. Distribución Normal

Estimamos los parámetros con la funcionalidad provista por Matlab

```
>>pd = fitdist(A, 'normal')
```

Siendo A el archivo de muestras. La estimación resultó

Tabla 1. Estimación de parámetros para una distribución normal

μ	σ
782.031 [769.863, 794.199]	697.628 [689.13, 706.34]

A continuación graficamos la distribución Normal(curva) comparada con el histograma(barras), Figura 1. La función de distribución acumulativa vs. La empírica, Figura2. El Q-Q Plot, y P-P Plot Figura 3,.

De los gráficos se desprende que asumir una distribución Normal para las muestras sería estar muy lejos de la realidad.

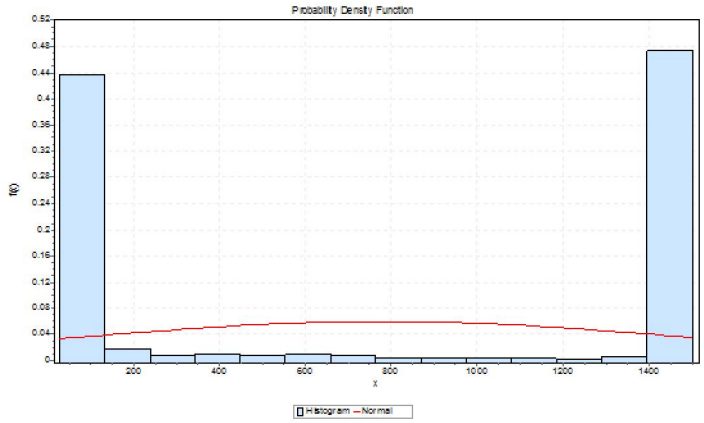

Fig. 1. Histograma y distribución Normal

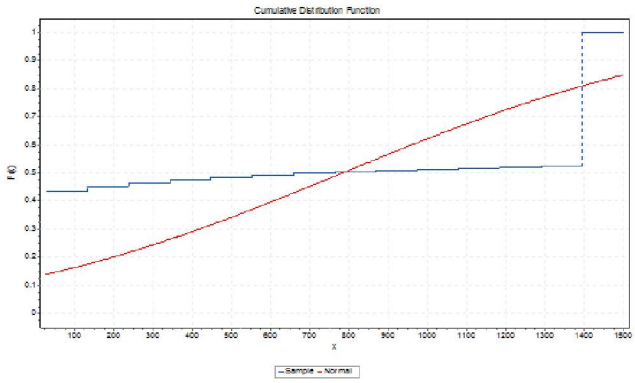

Fig. 2. Función de disttribución acumulativa vs. Empírica

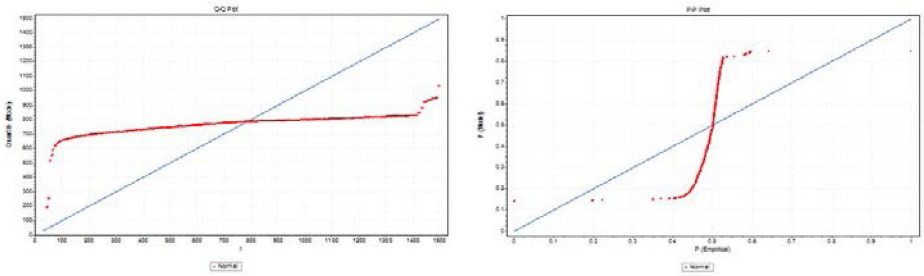

Fig. 3. Q-Q Plot y P-P Plot Muestras vs. Normal

En la Tabla 2 vemos los resultados de la Bondad de Ajuste

Tabla 2. Bondad de Ajuste para distribución Normal

Kolmogorov-Smirnov					
Sample Size	12630				
Statistic	0.29269				
P-Value	0				
α	0.2	0.1	0.05	0.02	0.01
Critical Value	0.00955	0.01088	0.01208	0.01351	0.0145
Anderson-Darling					
Sample Size	12630				
Statistic	1828.2				
α	0.2	0.1	0.05	0.02	0.01
Critical Value	1.3749	1.9286	2.5018	3.2892	3.9074
Chi-Squared					
Deg. of freedom	13				
Statistic	14714.0				
P-Value	0				
Rank	24				
α	0.2	0.1	0.05	0.02	0.01
Critical Value	16.985	19.812	22.362	25.472	27.688

4.2. Distribución Beta

Estimación de parámetros

Tabla 3. Estimación de parámetros distribución Beta

α_1	α_2	a	b
0.0707	0.05544	26.769	1500.0

Presentamos las mismas figuras que para el caso de la distribución Normal para terminar con la Bondad de Ajuste correspondiente.

Del análisis resulta que esta distribución resulta más realista, los puntos del Q-Q-Plot se aproximan a la recta de 45°.

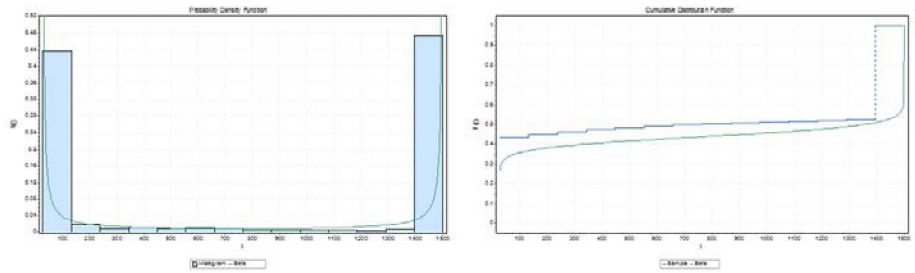

Fig. 4. Histograma- Dist.Beta(izq.) y Distribución Beta acumulativa-Empírica(der.)

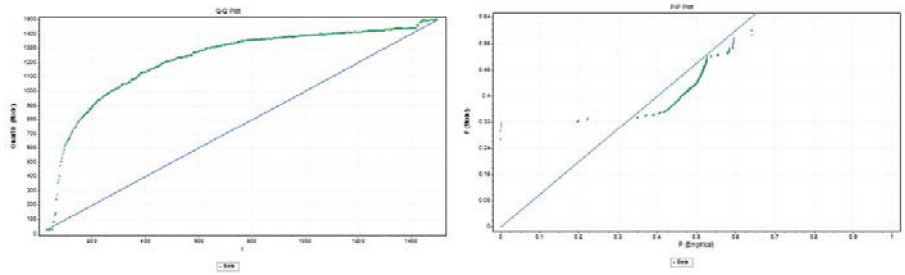

Fig. 5. Q-Q-Plot(izq.) y P-P Plot(der.) para la distribución Beta

Tabla 4. Bondad de Ajuste para distribución Beta

Beta					
Kolmogorov-Smirnov					
Sample Size	12630				
Statistic	0.31512				
P-Value	0				
α	0.2	0.1	0.05	0.02	0.01
Critical Value	0.00955	0.01088	0.01208	0.01351	0.0145
Anderson-Darling					
Sample Size	12630				
Statistic	13743.0				

α	0.2	0.1	0.05	0.02	0.01
Critical Value	1.3749	1.9286	2.5018	3.2892	3.9074

4.3. Distribución Johnson SB

La estimación resultó:

Tabla 5. Estimación de parámetros distribución

γ	δ	λ	ς
-0.01187	0.04707	1450.7	49.841

Las figuras y tabla de ajuste resultantes indican la mejor correspondencia de las muestras con esta distribución.

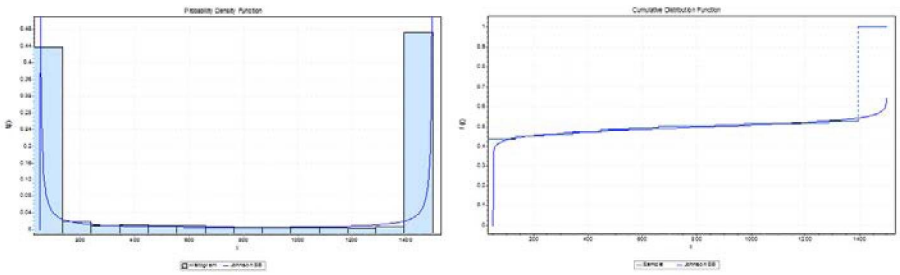

Fig.6. Histograma-Distribución(izq.),CDF-Empírica (der.) JohnsonSB

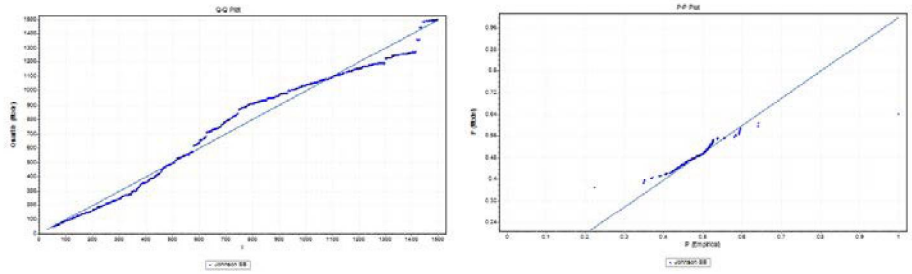

Fig. 7. . Q-Q-Plot(izq.) y P-P Plot(der.) para la distribución JohnsonSB

Tabla 6. Bondad de Ajuste para distribución JohnsonSB

Johnson SB					
Kolmogorov-Smirnov					
Sample Size	12630				
Statistic	0.35916				
P-Value	0				
α	0.2	0.1	0.05	0.02	0.01
Critical Value	0.00955	0.01088	0.01208	0.01351	0.0145
Anderson-Darling					
Sample Size	12630				
Statistic	9358.0				
α	0.2	0.1	0.05	0.02	0.01
Critical Value	1.3749	1.9286	2.5018	3.2892	3.9074

5. Conclusiones

En el presente trabajo hemos puesto en evidencia el recorrido necesario a cumplimentar para poder obtener una caracterización de tráfico que permita construir un modelo válidodel mismo. Entendemos por modelo válido aquel que nos permita diagnosticar problemas de performance en la red y/o aquel que nos permita predecir comportamiento y definir recursos necesarios en la implementación de una red de datos. Particularmente hemos apuntado a una etapa crítica del recorrido que es la elección de la distribución de la variable aleatoria objeto de análisis y que necesitamos definir para construir el modelo. Finalmente resaltamos la potencialidad de la distribución de JohnsonSB como elección válida. Sin duda no es una elección de carácter general, (con el tráfico de datos no es dable esa generalidad), pero si a tener en cuenta cuando tenemos una variable con dos valores de mayor recurrencia como fue el caso analizado. Somos concientes de que el trabajo no termina en esa elección. Sobre todo desde hace ya algo más de una década el tráfico ha cambiado su comportamiento, presentando un

caracter autosimilar del cual no hemos abordado su tratamiento por cuanto está fuera de los objetivos planteados del trabajo.

6. Referencias

[1] Hill, I.D., Hill, R., and Holder, R.L. (1976). Fitting Johnson curves by moments. *Applied Statistics.* AS99.

[2] Johnson, N.L. (1949). Systems of frequency curves generated by methods of translation. *Biometrika,* **36.** 149-176.

[3] Wheeler, R.E. (1980). Quantile estimators of Johnson curve parameters. *Biometrika.* **67-3** 725-728

[4] Karl Pearson (1905) Das Fehlergesetz und seine Verallgemeinerungen durch Fechner und Pearson. A Rejoinder, *Biometrika*, **4**, 169-212,

[5] Y.-N. Lee, X. Zhou, et. al (1998), Atmospheric chemistry and distribution of formaldhyde and several multioxygenated carbonyl compounds during the 1995 Nashvill/Middle Tennessee Ozone Study, *Journal of Geophysical Research, Vol 103, Issue D17:*22449-22462.

[6] F. George, K.M.Ramachandran (2009), *Analysis of Microarray Data for Gene Selection*, 25th Southern Biomedical Engineering Conference 2009; 15-17 May:237-238

[7] Lu, Y., O. A. Ramirez, R. M. Rejesus, T. O. Knight, and B. J. Sherrick. 2008. Empirically evaluating the flexibility of the Johnson family of distributions: a crop insurance application. Agricultural *& Resource Economics Review 37(1*): 79-91.

[8] Alexopoulos, C., D. Goldsman, J. Fontanesi, D. Kopald, and J. R. Wilson. (2008). Modeling patient arrivals in community clinics. *Omega 36*: 33-43

[9] Edward Prince (2012). *Mathematical Techniques in Crystallography and Material Science,* Springer Science & Business Media.

[10] Fonseca, T.F., Marques, C.P., Parresol, B.R(2009).: *Describing maritime pine diameter distributions with Johnson's S B distribution using a new all-parameter recovery approach.* For. Sci. 55(4), 367–373 (2009)

[11] Ayana Mateus, Margarida Tomé (2013). *Fitting Johnson's SB Distribution to Forest Tree Diameter.* Advances in Regression, Survival Analysis, Extreme Values, Markov Processes and Other Statistical Applications pp 289-296. Springer

[12] Jones, D. L. (2014). *Johnson Curve Toolbox for Matlab: analysis of non-normal data using the Johnson family of distributions.* College of Marine Science, University of South Florida, St. Petersburg, Florida, USA