Predicción o pronóstico. Caso para un sistema complejo

José Cruz Ramos Báez, jcramos@up.edu.mx
María de Lourdes Martínez Villaseñor, lmartine@up.edu.mx
Lorenzo Miguel Elguea Fernández, lelguea@up.edu.mx

RESUMEN

El tema de predicción o pronóstico aparece en todo momento al solucionar problemas de toma de decisiones, sobre todo al utilizar grandes bases de datos. En este artículo se comentan métodos estadísticos para toma de decisiones o inferencias, primero empleando la probabilidad de un evento, segundo usando regresión lineal, finalmente como objetivo principal, se emplea una base de datos de un sistema complejo o Distribuidor de Señales Digitales (DSD) y con ayuda de las herramientas de aprendizaje máquina predecir una falla del sistema.

Palabras clave: Predicción, pronóstico, regresión, aprendizaje máquina, sistema complejo.

PREDICTION OR FORECAST CASE FOR A COMPLEX SYSTEM

ABSTRACT

Prediction or forecast appears frequently to solve making decisions problems, especially when using large databases. Statistical methods to support decisions are discussed, first determining the probability of an event, second using linear regression, finally main objective in this paper we use machine learning techniques for failure prediction in a Digital Signal Distribution complex system.

Key words: Prediction, forecast, regression, Machine Learning, complex system.

INTRODUCCIÓN

Evitar el mal funcionamiento o la interrupción de los sistemas de comunicación y la maquinaría así como la sinergia de ambas implican grandes tiempos de monitoreo, metodologías de administración con tiempos base para mantenimiento y diagnóstico de los dispositivos mecánicos y eléctricos, pero los sistemas computacionales necesitan de estos y otros métodos para evitar un mal funcionamiento o detención de los sistemas (falla) [1].

Los sistemas complejos como el de comunicación y de internet están compuestos de varios elementos, componentes y de software para su funcionamiento, incluyen sistemas de distribución de señales digitales (DSD) y otros, necesitan de métodos de mantenimiento que eviten fallas en los sistemas y componentes dados los altos costos de una falla o detención del sistema complejo.

Existen distintos métodos para pronosticar fallos en componentes [1], algunos estadísticos y otros de aprendizaje máquina; sin embargo, actualmente, la detección de fallos en un sistema complejo es un tema de investigación[2].

El objetivo principal de este artículo es mostrar, con la ayuda de una base de datos de un sistema complejo o distribuidor de señales digitales (DSD), un método para inferir fallas en el DSD utilizando métodos estadísticos y herramientas de aprendizaje máquina [3].

Se mostrará con tres experimentos algunas de las herramientas estadísticas para inferir y en su caso predecir resultados, a su vez entender que al aumentar el número de variables complica el inferir y podemos auxiliarnos de las herramientas de aprendizaje máquina.

De manera muy breve se comentan los experimentos que posteriormente se tratarán con mayor extensión.

Es muy común cometer el error de suponer o confundir las palabras predicción y pronóstico como sinónimos, esto implica conocer la diferencia entre ambas, cuyas definiciones son:

- Predicción[1]: acción y efecto de anunciar un hecho antes de que ocurra o que se producirá en el futuro.
- Pronóstico: es una afirmación sobre un evento cuya ocurrencia no es segura.

Existe una gran diferencia entre ambas definiciones, pero la predicción en una empresa tanto para producción, fuerza laboral o costos implica grandes decisiones.

El cálculo de la probabilidad de un evento es un porcentaje con el cual podemos inferir la ocurrencia del mismo [4]. El primer experimento probará esta parte.

Una herramienta estadística muy empleada en la predicción es la regresión lineal.

La regresión lineal relaciona dos variables, aunque deben hacerse pruebas que confirmen el uso de la regresión, entre estas pruebas está el Coeficiente de Determinación R^2 [4, 5]. En el segundo experimento se verá la aplicación de la regresión lineal. En el tercer experimento se obtiene una predicción con las herramientas de aprendizaje máquina.

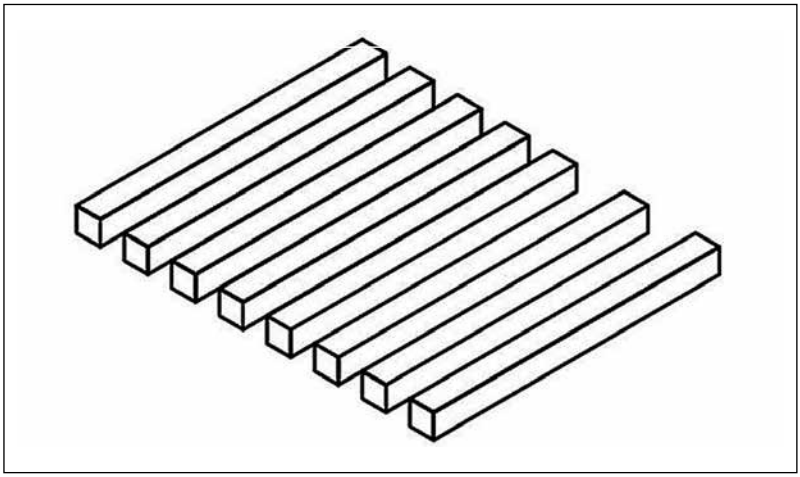
[1] http://dem.colmex.mx/moduls/Default.aspx?id=8. Fecha de acceso: septiembre 16, 2016.

MÉTODOS DE PREDICCIÓN Y PRONÓSTICO

Experimento 1
Probabilidad

Otro tema relevante de los datos estadísticos es hacer inferencias de los mismos, haciendo un ejemplo sencillo para convencernos de lo que involucra una inferencia sobre los datos estadísticos realizamos un experimento usando una figura imposible mostrada en la Figura 1 y dando un límite de tiempo (1 min) pedimos responder a la pregunta ¿cuántas barras completas observas?[2]

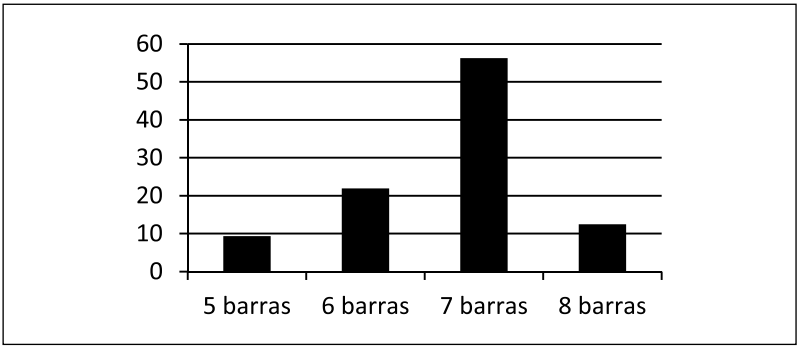
Figura 1. *Figura imposible.*



Realizando este experimento con varias personas obtenemos la gráfica de respuestas mostrada en la Figura 2.

[2] http://www.dailymail.co.uk/femail/article-3695569/Can-count-bars-Seemingly-simple-optical-illusion-brain-doing-somersaults-sweeps-web.html. Fecha de acceso: agosto 16, 2016

Figura 2. *Respuestas al número de barras en la figura imposible*



Observamos que la mayoría, un 56.25% de encuestados dan una respuesta de 7 barras, sin embargo, quienes respondieron correctamente fue el 21.88% (6 barras) dando un total de 78.12% de personas que ¡cometieron un error en el conteo![3]

Algunas ideas a inferir son: fueron poco observadores, no recuerdan cómo contar, es imposible hacerlo en poco tiempo, la figura no es muy visible, etcétera. No es posible dar respuesta a alguna de estas inferencias con este simple experimento, pero una inferencia general puede ser "existe una alta probabilidad de que una persona se equivoque en la respuesta" y aún para probar este resultado se requiere un mayor análisis, por ejemplo ¿dependerá de la edad? Con este análisis nos damos cuenta de lo relevante que es hacer inferencias con los resultados estadísticos, esto implica considerar más tiempo de lo que uno puede imaginar o programar.

[3] Experimento realizado en auditorio de la UP Mixcoac, el 29 de septiembre 2016.

Experimento 2
Regresión Lineal

Otra técnica estadística para hacer un pronóstico es la Regresión Lineal, la cual relaciona dos variables, una dependiente y la otra independiente, determinando si tienen correlación se ajustan a una línea recta.

El método empleado para obtener la línea recta es el de mínimos cuadrados ordinarios (MCO), este método puede ser muy determinista.

Con este procedimiento podemos determinar el valor de la variable dependiente para un cierto valor de la variable independiente. Para validar este resultado se emplea el Coeficiente de Determinación R^2, indica el porcentaje de variabilidad conque la variable dependiente es explicada o descrita por la recta de regresión, mientras mayor es el porcentaje mejor se ajustan los puntos a la recta de regresión.

Tabla 1. *Precio de construcción por área construida.*

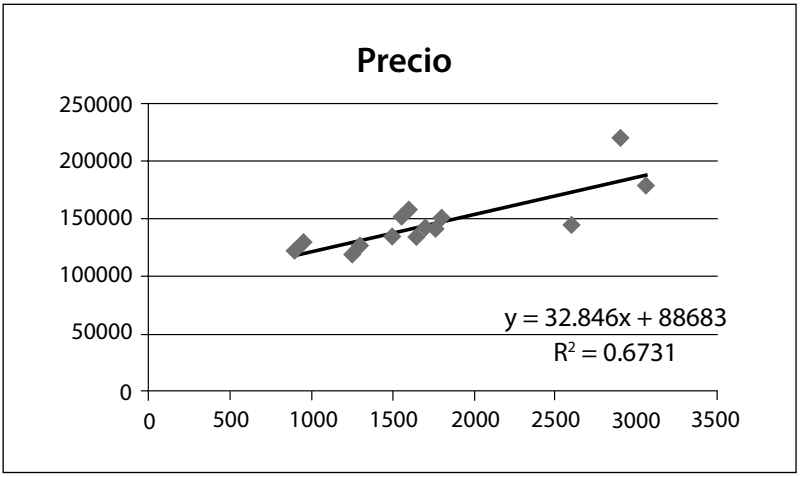
Área(m^2)	Precio
900	122500
950	129000
1250	118500
1300	125000
1300	126500
1500	135000
1550	152000
1600	158000
1650	134500
1700	142000
1750	141000
1800	150000
2600	145000
2900	220000
3060	179000

La tabla 1 muestra los precios de construcciones por área construida.

Al poner los puntos en el eje (diagrama de dispersión) se observa la relación de a mayor área mayor precio y podemos suponer una relación lineal lo que nos lleva a determinar una recta de regresión. Utilizando una hoja de cálculo en Excel se obtiene la

Figura 3 calculando la ecuación de la recta de ajuste y el coeficiente de determinación.

Figura 3. *Recta de ajuste precio vs área.*



En el cálculo $R^2 = 0.6731$ indicando que la variable dependiente (precio) queda determinada un 67.31% por la recta.

Es un porcentaje relativamente alto y para un mejor ajuste determinamos un intervalo de confianza al 95%, esto indica que el valor de la variable dependiente quedará dentro de ese intervalo de acuerdo a ese porcentaje.

El intervalo de confianza (Figura 4) queda descrito por dos rectas paralelas a la recta inicial y no están muy separadas, siendo una característica de un buen ajuste.

En este análisis también podemos encontrar los intervalos de predicción, por lo general son líneas curvas abriéndose. Mientras se aleje del dominio de la recta de regresión, los intervalos son mayores perdiendo exactitud en predicciones muy lejanas. La Figura 5 muestra dichos intervalos.

Figura 4. *Intervalo de confianza de la regresión.*

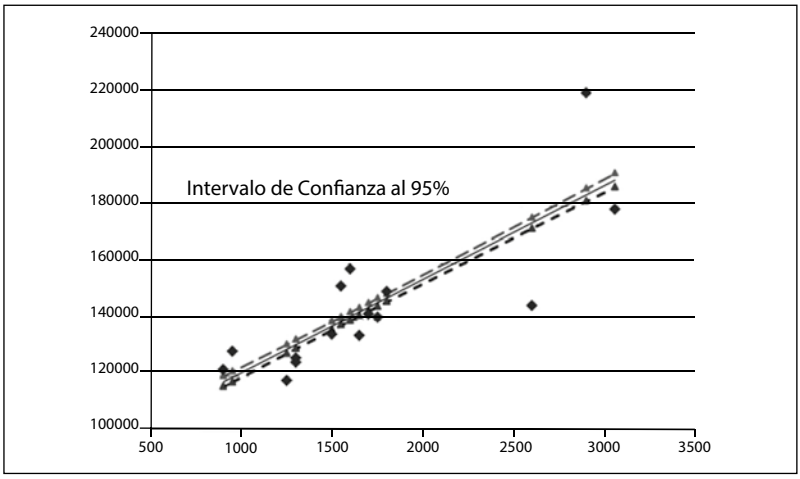
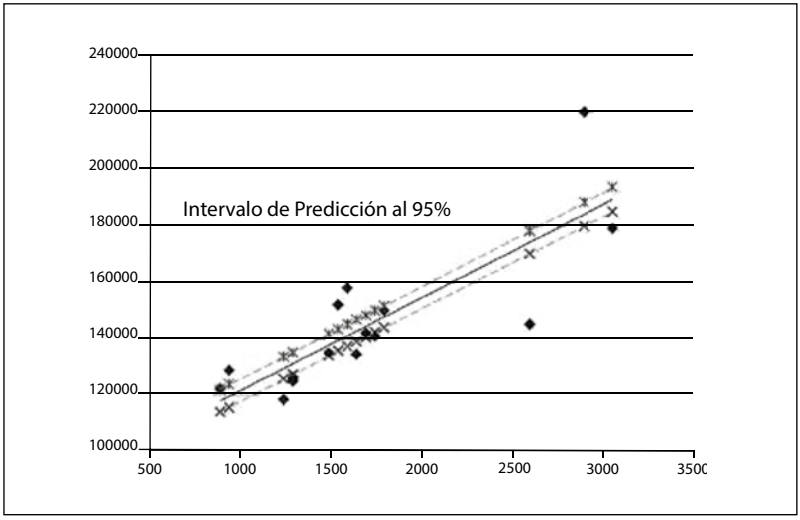


Figura 5. *Intervalo de predicción de la regresión.*



Observemos cómo son más abiertos que los intervalos de confianza.

Pronosticar un valor de una construcción con un área mayor a 3060 m^2 implica que no debe alejarse mucho de este valor para no incrementar los intervalos de predicción.

Calculando este intervalo de predicción tendremos el valor de la construcción entre dos valores y no un valor único o determinado. Esto implica hacer un pronóstico muy confiable [4] con el fin de no tener un intervalo muy grande.

Experimento 3
Aprendizaje Máquina

En los últimos años los procesos predictivos se vienen realizando con herramientas de aprendizaje máquina y otros más, entre estas herramientas están: regresiones, agrupamientos, árboles de decisión, redes neuronales, algoritmos genéticos, máquinas de soporte vectorial, y otras más.

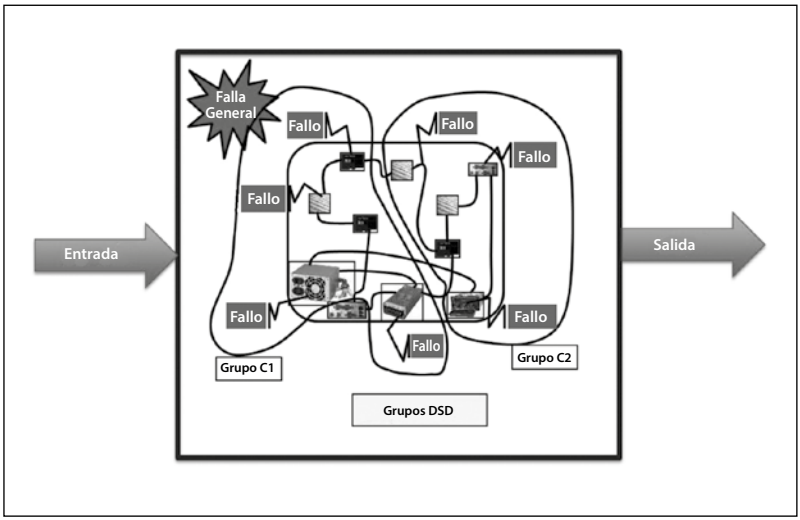
Estas herramientas ayudan al análisis de grandes bases de datos y también apoyándose en minería de datos se buscan relaciones, patrones, inferencias, etcétera sobre los resultados. Existen programas que contienen varias de estas herramientas, en particular se utilizó WEKA [5] para este artículo.

La aplicación de estas herramientas será sobre un sistema complejo.

Un sistema complejo consta de un gran número de componentes de las cuales se desea que no tengan alguna falla, el sistema puede ser mecánico o un sistema de distribución de señales digitales (DSD) para internet y telefonía celular, (Figura 6).

Para ello el sistema se toma como una caja negra donde se tiene una señal de entrada y, después de decodificarla, una de salida.

Figura 6. *Esquema DSD.*



Para el experimento los datos se obtuvieron de un DSD privado. Analizando los datos de fallos registrados y buscando aquellos que provocan la falla total del sistema.

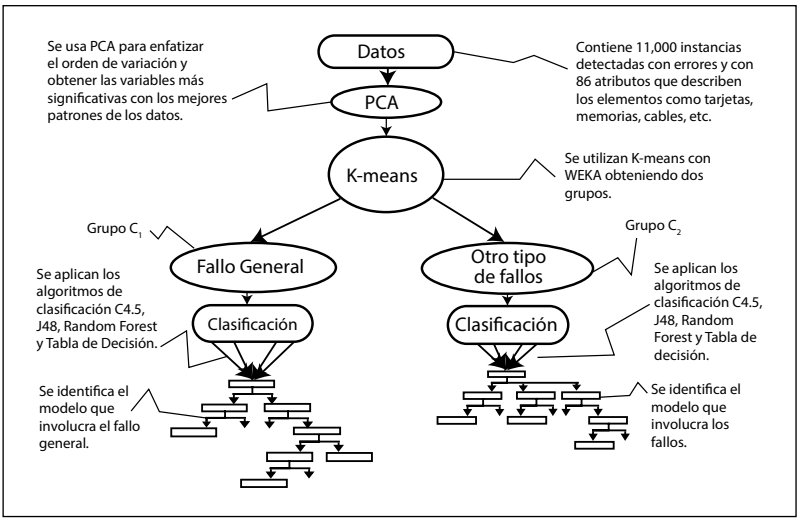
Predicción para el caso de un DSD

- **Objetivo:** Mediante una base de datos, determinar en un sistema complejo los posibles fallos que provocarán una falla del sistema.

- **Desarrollo:** La base tiene aproximadamente 11000 datos, por lo cual primero se revisan todas las correlaciones entre características, determinando las mejores de ellas con el análisis de componentes principales (PCA), posteriormente se buscan grupos o clasificaciones, encontrando dos grupos característicos (Fig. 6), donde en uno de ellos que-

daron las fallas del sistema, posteriormente con un árbol de decisión se encuentran las líneas que llevan a la falla del sistema [3]. La Fig. 7 muestra un esquema de los resultados y métodos utilizados.

Figura 7. *Diagrama de solución a fallas en el DSD.*



CONCLUSIONES

Durante varios años el uso de la estadística en la toma de decisiones se aplicó con buenas aproximaciones y cuando los recursos computacionales aumentaron esta aproximación tuvo mayor exactitud, en estos días con las herramientas de aprendizaje máquina se tiene una herramienta más poderosa para el manejo de grandes cantidades de datos.

Estas herramientas no dejan de utilizar procesos estadísticos, sin embargo, los cálculos complicados se realizan con gran rapidez permitiendo utilizar mayor tiempo en el análisis.

No importa el modelo estadístico que utilicemos para probar una hipótesis lo que se debe tomar en cuenta es todo lo necesario para su análisis.

Las herramientas de aprendizaje máquina son más poderosas y permiten revisar varios métodos para obtener relaciones y patrones no observables, y con ello predecir comportamientos futuros.

El uso de grandes bases de datos implica tiempo en comprender cada característica y algunos expertos mencionan que por lo menos el 70% del tiempo empleado para predecir un comportamiento es utilizado simplemente para entender la base de datos.

Las herramientas de aprendizaje máquina iniciaron su uso en el área de ciencias e ingeniería, pero a últimas fechas se emplean en finanzas[6], economía [7] y otras áreas, sin duda pueden emplearse en Hospitalidad y Recursos Humanos puesto que existen grandes bases de datos en estas áreas.

FUENTES CONSULTADAS

1. Salfner, F., Lenk,M., and Miroslaw Malek: A survey of online failure prediction methods. ACM Comput. Surv. vol. 42, no. 3, Article 10 (March 2010), 42 pages. DOI=10.1145/1670679.1670680 http://doi.acm.org/10.1145/1670679.1670680.
2. Zhiling Lan, Jiexing Gu, Ziming Zheng ; A Study of Dynamic Meta-Learning for Failure Prediction in Large-Scale Systems; Journal of parallel and distributed computing; vol. 70, no6, pp. 630-643, 2010
3. Ramos-Báez, José Cruz; Martínez-Villaseñor, María de Lourdes; Rosso-Pelayo, Dafne (2015); Methodology model for failure prediction in a Digital Signal Distribution; Research in Computing Science 104 (2015); pp. 91–101.

4. Lind, D; Marchal, W; Wathen, S; Estadística aplicada a los negocios y la economía; 15ª edición, 2008, ed Mc Graw Hill.
5. Hanke, John, Wichern, Dean (2010); Pronósticos en los Negocios; México, 9ª Edición, Ed Pearson.
6. Molina López, José Manuel, García Herrero, Jesús (2012); Técnicas de Análisis de Datos, Aplicaciones Prácticas Utilizando Microsoft Excel y WEKA; Libro, Universidad Carlos III de Madrid, España, http://ocw.uc3m.es/ingenieria-informatica/analisis-de-datos/libroDataMiningv5.pdf
7. Sneha Soni (2010); Applications of ANNs in Stock Market Prediction: A Survey; International Journal of Computer Science & Engineering Technology (IJCSET); Vol. 2 No. 3; ISSN : 2229-3345
8. Cabrera Llanos, Agustín I.; Ortiz Arango, Francisco (2011); Pronóstico del rendimiento del IPC (Índice de Precios y Cotizaciones) mediante el uso de redes neuronales diferenciales; Contaduría y Administración, IPN, Vol. 57, No. 2, abril-junio 2012: 63-81.

Referencias WEB

http://dem.colmex.mx/moduls/Default.aspx?id=8
http://www.dailymail.co.uk/femail/article-3695569/Can-count-bars-Seemingly-simple-optical-illusion-brain-doing-somersaults-sweeps-web.html
http://ocw.uc3m.es/ingenieria-informatica/analisis-de-datos/libroDataMiningv5.pdf