Aplicación de los algoritmos genéticos para seleccionar el mejor modelo de regresión de Cox

Application of genetics algorithms for selecting the best regression Cox model

Douglas Rivas[1*], Luciano Maldonado[2†],Rafael Borges[2‡]

[1]Instituto Universitario Tecnológico de Ejido, Mérida

[2]Universidad de Los Andes, Mérida

Resumen

Este trabajo, enmarcado en el campo de la Computación Evolutiva, presenta el desarrollo de un Algoritmo Genético para encontrar los parámetros óptimos de un Modelo de Regresión de Cox del Análisis de Supervivencia para pacientes del servicio de Diálisis Peritoneal del Hospital Clínico Universitario de Caracas entre 1980 y 2000, (Borges 2002). Se hace uso de la técnica de los Algoritmos Genéticos como método de búsqueda de una mejor estimación de los parámetros del Modelo de Cox al obtenido por los métodos clásicos de optimización. El algoritmo fue programado completamente en el lenguaje C++, bajo un diseño de programación modular. Las características principales del algoritmo son: a) La población inicial, que está constituida por 10 individuos; se genera de manera aleatoria entre un rango de valores; dicho rango fue obtenido luego de realizar diversas pruebas; b) la función de ajuste se basó en el Criterio de Información de Akaike (AIC); c) la selección de los individuos a reproducirse se realizó por torneo; e) para el cruce se usó el operador multipunto y la mutación se realizó a todos los genes de una parte de los cromosomas de la población. El algoritmo desarrollado permitió obtener estimaciones de los parámetros del Modelo de Regresión de Cox, y con mejor valor de AIC, a los obtenidos utilizando los métodos clásicos.

Palabras clave: Algoritmo Genético, Modelo de Regresión de Cox, Criterio de Información de Akaike (AIC), Análisis de Supervivencia.

Abstract

This work, involved in the area of evolutive computing, presents the development of a Genetic Algorithm in finding the optimal parameters of a Cox Regression Model for patients of the Service of Peritoneal Dialysis of the "Hospital Clínico Universitario de Caracas"between 1980 and 2002, performed by (Borges 2002). The technique of the genetic algorithms is used as a method for finding a better estimation of the parameters of the Cox Model than the obtained by the classical optimization methods. The algorithm was completed programmed in the language C++, using a modular programming design. The main characteristics of the algorithm are: a) The initial population, of 10 subjects, is generated randomly between a range of values, this range was obtained after several essays, b) The adjustment function was based in the Akaike Information Criteria (AIC), c) The selection of the subjects to be reproduced was done by tournament, d) The multipoint operator for the crossing and, the mutation was done to all the genes of one part of the chromosomes of the population. The developed algorithm was useful to obtain the estimation of the Cox Regression Model and with better AIC values than the obtained by the classical methods.

Keywords: Genetic Algorithms, Cox Regression Model, Akaike Information Criteria (AIC), Survival Analysis.

*Profesor. Email:drivas@ula.ve
†Profesor. Email:maldonaj@ula.ve
‡Profesor. Email:borgesr@ula.ve

1. Introducción

1.1. Formulación del Problema

El análisis de supervivencia permite estudiar y construir modelos para analizar el tiempo que un suceso tarda en ocurrir. En dicho análisis o proceso las diferentes variables pronóstico permiten estimar el tiempo de aparición del suceso. Entre los diferentes tipos de modelos que se pueden emplear, uno de los más extendidos es el modelo de riesgos proporcionales, también conocido como Modelo de Cox. Este, al igual que los demás modelos estadísticos paramétricos, enfrenta el problema de estimar sus parámetros a partir de datos observados. Es bien conocido que en todos los problemas de estimación la solución no es única; por el contrario, se tienen varias alternativas que funcionan como mínimos o máximos locales de cierta función. Por lo tanto, los problemas de estimación de los parámetros de los modelos de regresión no es otra cosa que un problema de la búsqueda de aquellos parámetros que representen mejor los datos en estudio.

Para resolver distintos problemas, en los últimos años ha tomado auge una técnica de la computación evolutiva conocida como algoritmos genéticos. Los algoritmos genéticos son mecanismos de optimización de funciones basados en leyes biológicas de selección natural que presentan ventajas con respecto a los métodos tradicionales de optimización. En especial, estos métodos iterativos convergen al óptimo global de la función objetivo de análisis, sin importar su grado de complejidad y su dominio, evitando procedimientos adicionales que en ocasiones se desvían hacia óptimos locales.

Este trabajo propone un método alternativo para estimar los parámetros del Modelo de Regresión de Cox haciendo uso de un Algoritmo Genético. Específicamente, el algoritmo calcula los valores óptimos de los parámetros del Modelo de Regresión de Cox en el Análisis de Supervivencia para pacientes del servicio de Diálisis Peritoneal del Hospital Clínico Universitario de Caracas entre 1980 y 2000, (Borges 2002).

1.2. Justificación: importancia, aplicaciones esperadas

El análisis de sobrevivencia comprende un conjunto de técnicas de gran importancia en diversas aplicaciones prácticas, pues permite estudiar la variable tiempo hasta que ocurre un evento y su dependencia de otras posibles variables explicatorias. Debido a su amplia aplicación y muy especialmente al amplio uso del modelo de regresión de Cox como técnica para estimar la variable tiempo en función de diversas variables explicativas, resulta interesante construir un algoritmo genético, el cual pueda seleccionar el mejor modelo de Cox para una serie de datos. Un algoritmo de tales características resultaría en una herramienta de gran utilidad en la medicina, la ingeniería y en la estadística, pues además de presentarse como un método alternativo que selecciona los valores óptimos de los parámetros de un Modelo de Regresión de Cox, el éxito del mismo motivará futuras aplicaciones de los Algoritmos Genéticos en el campo de la estadística.

2. Metodología de la investigación

Para el desarrollo de este proyecto se sigue una metodología basada principalmente en dos aspectos. En primer lugar, la revisión exhaustiva de los fundamentos teóricos de los Algoritmos Genéticos y del Análisis de Supervivencia, específicamente del Modelo de Regresión de Cox; y en segundo lugar, la definición de las actividades inherentes a la construcción y pruebas del Algoritmo.

3. Análisis del Caso de Estudio

El Algoritmo Genético que se construye en esta investigación selecciona el mejor Modelo de Regresión de Cox del Análisis de Supervivencia para pacientes del servicio de Diálisis Peritoneal del Hospital Clínico Universitario de Caracas entre 1.980 y 2.000.

Para dicho análisis se registraron un conjunto de variables para 246 individuos en Diálisis Peritoneal, que ingresaron al servicio entre el 2 de Junio de 1980 y el 6 de Diciembre de 1996, y salieron entre el 15 de enero de 1981 y el 31 de octubre de 2000.

Luego de realizar un estudio exhaustivo, probando la significancia de cada una de las variables, se incluyeron las siguientes variables significativas al 10 %: Diabetes, edad y quetellet; lo cual se puede apreciar en la tabla 1.

Tabla 1: Estimación de los coeficientes para el modelo definitivo de Cox para los datos según meses.

Covariable	coef	exp(coef)	ee(coef)	Z	p
diabetes	0.5492	1.732	0.3208	1.71	0.087
edad	0.0315	1.032	0.0097	3.25	0.0011
quetellet	-0.0969	0.908	0.0389	-2.49	0.013

Donde:

coef es el coeficiente estimado mediante el modelo.

exp(coef) es el exponencial del coeficiente y se interpreta como el riesgo.

ee(coef) es el error estándar del coeficiente.

Z es el estadístico de contraste para la significación del coeficiente.

P es el p-valor de probabilidad de la significación del coeficiente.

Por lo tanto el Modelo de Regresión de Cox tiene la siguiente forma:

$$\lambda(t, Z(t)) = \lambda_0(t)e^{\beta_1 Z_1(t)+\beta_2 Z_2(t)+\beta_3 Z_3(t)} \tag{1}$$

Donde

β_1, es el parámetro asociado con la variable diabetes.

β_2, es el parámetro asociado con la variable edad.

β_3, es el parámetro asociado con la variable quetellet.

Los riesgos asociados a cada uno de los parámetros y sus respectivos intervalos de confianza al 95 % se muestran en la tabla 2.

Tabla 2: Exponencial de los coeficientes para el modelo definitivo de Cox para los datos según meses.

Covariable	Exp(coef)	Exp(-coef)	LCI(95)	LCS(95)
diabetes	1.732	0.577	0.924	3.248
edad	1.032	0.969	1.013	1.052
quetellet	0.908	1.102	0.841	0.979

4. El Algoritmo Genético

A continuación se establecen cada uno de los pasos realizados para la construcción del Algoritmo.

4.1. Elementos del Algoritmo Genético

A continuación se describen cada uno de los elementos usados en la construcción del algoritmo basados en la teoría anteriormente señalada:

4.1.1. Selección de la variable de entrada

En este caso la variable de entrada o cromosoma está conformado por los tres parámetros del Modelo de Regresión de Cox; es decir, los parámetros de las variables diabetes (β_1), edad (β_2) y quetellet (β_3). Por lo tanto, el cromosoma es un vector fila con tres elementos:

$$cromosoma = \left[\beta_1, \beta_2, \beta_3\right]$$

cuyo rango de valores está entre -5 y 5. Los Valores fuera de este rango hacen que la función de ajuste tienda a infinito.

4.1.2. Representación o codificación de la variable de entrada

Aunque el rango de valores de los parámetros es un subconjunto de los números reales, las soluciones se codificaron como cadenas binarias de secuencias de 1s y 0s, cuya longitud depende del valor del parámetro más grande en valor absoluto.

4.1.3. Tamaño de la Población

Como se establece en la teoría, no existe un método para determinar el número óptimo de individuos, en este caso, se realizaron diversas corridas con diferentes tamaños de población inicial, obteniéndose que 10 individuos eran suficientes para realizar el estudio, ya que, un número mayor de individuos no modificaba los resultados obtenidos con 10 individuos.

4.1.4. Elección de la población inicial

La población inicial se generó aleatoriamente entre el rango de valores de los parámetros anteriormente fijados.

4.1.5. Función de aptitud o de adaptación

La función de aptitud utilizada está basada en el criterio AIC, el cual es la maximización del logaritmo de la verosimilitud de la función de riesgo proporcional de Cox. Por lo tanto, se usó como función de ajuste la siguiente función:

$$AIC = -2\log[L(\beta)] + \alpha \tag{2}$$

Donde

$$L(\beta) = \prod_{i=1}^{k} \frac{exp\{Z_i(t_i)\beta\}}{\sum_{l \in R(t_i)} exp\{Z_l(t_i)\beta\}} \tag{3}$$

$t_1 \leq t_2 \leq ... \leq t_k$ son los tiempos de fallas observados, asumidos distintos en la muestra; $Z_i(t_i)$ es el vector de covariables en el tiempo t_i para el sujeto que falla en el tiempo t_i ; y $Z_l(t_i)$ es el vector de covariables correspondientes para el l-esimo miembro de $R(t_i)$, el conjunto de individuos en riesgo en el tiempo t_i, (Cox & Oakes 1984).

4.1.6. Selección

Una vez evaluados los individuos de la población, se usó el método de selección por torneos; es decir, se seleccionaban la mitad de los individuos que tenían mayor porcentaje de contribución a la función de ajuste. En este caso, como la función de ajuste está basada en el criterio AIC, se seleccionaron los individuos que tenían menor AIC.

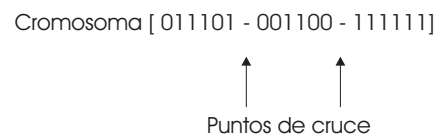
4.1.7. Operadores

Se usaron los dos operadores genéticos principales: cruce y mutación.

Cruce. Para el operador de cruce se tomaron en cuenta dos elementos: la probabilidad de que ocurriera el cruce y el número de puntos para realizar el cruce. En cuanto a la probabilidad de cruce, se tomó 0.8 como probabilidad de cruce, ya que es común que este operador ocurra en un algoritmo genético. Para el número de puntos de cruce se realizó un multipunto de dos puntos. Es decir, supongamos el siguiente cromosoma:

$$cromosoma = [011101 - 001100 - 111111]$$

Entonces los dos puntos de cruce están distribuidos de la siguiente manera:



Mutación. Para la mutación se tomaron en cuenta tres elementos: la probabilidad de que ocurra la mutación, cuántos individuos mutar y cuáles elementos de cada individuo mutar. La probabilidad de mutación seleccionada fue 0.3 y en cuanto al número de elementos a mutar se realizó de forma aleatoria, tomando en cuenta que no se podían mutar todos los individuos y por lo tanto, el de menor AIC no se selecciona para la mutación. En lo que respecta a cuáles elementos de los individuos seleccionados anteriormente mutar, se usó el criterio de mutar todos los elementos de los individuos seleccionados.

4.1.8. Criterio de Parada

El criterio de Parada establecido se obtuvo por observación en las pruebas. Es decir, se realizaron varias pruebas con distintos número de iteraciones y se escogió aquel número de iteraciones donde todos los individuos tenían el mismo valor de AIC.

4.2. Diseño del Algoritmo Genético

Una vez establecidos los elementos que componen el Algoritmo Genético, se procedió a su diseño. Para ello, se tomó en cuenta el diseño de programación modular; es decir se crearon una serie de módulos que estaban encargados de realizar acciones muy específicas, que al interrelacionarlos, cumplieron con el propósito del proyecto en estudio. Entre los principales módulos creados se encuentran:

- **Generación de la población inicial.** Genera una población de n individuos, con tres parámetros, donde cada parámetro está en un rango de valores previamente establecidos.
- **Evaluación de la Función de Ajuste.** Evalúa la población en la función de ajuste o adaptación; es decir, calcula el valor AIC para cada uno de los individuos.
- **Ordenamiento.** Ordena los individuos de menor a mayor de acuerdo al valor AIC calculado anteriormente.
- **Selección y Reproducción.** Selecciona la mitad de los individuos con menor AIC y los reproduce eliminando a su vez la mitad con mayor AIC.
- **Conversión o codificación.** Convierte los parámetros de cada uno de los individuos en su respectiva representación binaria.
- **Creación del Cromosoma.** Crea los cromosomas compuestos por las cadenas binarias obtenidas anteriormente.
- **Cruce.** Realiza el cruce de los cromosomas obtenidos anteriormente.
- **Mutación.** Realiza la mutación de los individuos, asignando el valor 0 donde hay un 1 y viceversa.

Todos estos módulos, y otros no menos importantes, se interrelacionaron entre si tomando en cuenta el número de generaciones necesarias para obtener una población cuya función de ajuste convergiera en una única solución.

5. Pruebas y Resultados

Una vez construido el Algoritmo Genético y codificado en el lenguaje C++, se procedió a realizar diversas pruebas variando algunos parámetros, tales como el número de individuos, el rango de los parámetros del modelo y el número de iteraciones. Se obtuvo en cada uno de los casos, valores diferentes de los parámetros a estimar del Modelo de Regresión de Cox y valores diferentes del AIC.

En primer lugar, se fue variando el número de iteraciones. El método de variación fue el decreciente; es decir, se comenzó con 100 iteraciones y se fue disminuyendo de 10 en 10, comprobando que todos los valores del AIC convergieran a un único valor. Se obtuvo que para una población de 10 individuos era suficiente 30 iteraciones. Luego, se tomaron diferentes tamaños para la población inicial pero los resultados no cambiaban significativamente. Por lo tanto, se estableció como 30 el número de iteraciones y 10 el tamaño de la población. En cuanto al rango de valores sobre el cual estaban definidos los parámetros de la población inicial se eligió el intervalo [-5,5], valores fuera de este rango conducían a valores muy altos de AIC e incluso en algunos casos resultaba infinito.

Una vez establecidos estos elementos del algoritmo se realizaron diversas corridas. En la tabla 3 se muestran los resultados para 10 pruebas del algoritmo.

Tabla 3: Valores estimados de los parámetros del Modelo de Regresión de Cox con sus respectivos valores de AIC, en 10 pruebas del algoritmo.

Prueba	β_1	β_2	β_3	AIC
1	0.533051	0.070153	-0.584849	402.729
2	-0.052507	0.027918	-0.383819	407.276
3	-0.099685	0.04056	-0.485062	394.507
4	-0.502085	-0.022684	-0.353294	447.779
5	0.425578	-0.010896	-0.317306	430.429
6	-0.479252	0.034671	-0.520996	394.928
7	0.423431	-0.015299	-0.437233	421.427
8	-0.2721	-0.008445	-0.39926	424.742
9	-0.536755	0.05092	-0.441887	406.571
10	0.756753	-0.014112	-0.747774	428.597

Como puede observarse, los valores obtenidos de AIC en cada una de las corridas son menores que el obtenido por el método clásico (Borges 2002), ver tabla 4.

Tabla 4: Valores estimados de los parámetros del Modelo de Regresión de Cox con su respectivo valor de AIC, (Borges 2002)

β_1	β_2	β_3	AIC
0.5492	0.0315	-0.0969	507.542

Al comparar los resultados de la tabla 3 y la tabla 4 se observa que el algoritmo genético obtiene estimaciones de los parámetros con mejor AIC que el método clásico, ver figura 1, ya que todos los valores de AIC para las diez pruebas están por debajo del valor de AIC para el resultado obtenido con dicho método.

Por otro lado, se puede ver que los resultados obtenidos varían de prueba en prueba, esto implica que las estimaciones dependen de la población inicial que se genera aleatoriamente; por lo tanto, al igual que como ocurre con los métodos clásicos, el algoritmo desarrollado no da la seguridad de obtener un óptimo global, sin embargo da una mejor solución que el método clásico. Es importante resaltar que las estimaciones de los parámetros varían de prueba en prueba, y que en algunas de ellas estas estimaciones difieren significativamente de las obtenidas por el método clásico; aún así, el porcentaje de estas ocurrencias es bajo.

En la tabla 5, se muestra el riesgo asociado a cada uno de los parámetros estimados por el algoritmo.

Al comparar estos valores con los intervalos de confianza del riesgo asociado con los parámetros obtenidos por el método clásico (tabla 2), se observa que un alto porcentaje de los valores obtenidos por el algoritmo genético caen dentro del intervalo respectivo.

Todos estos resultados indican que el algoritmo genético desarrollado podría usarse como una técnica alternativa para estimar los parámetros del Modelo de Regresión de Cox del caso en estudio.

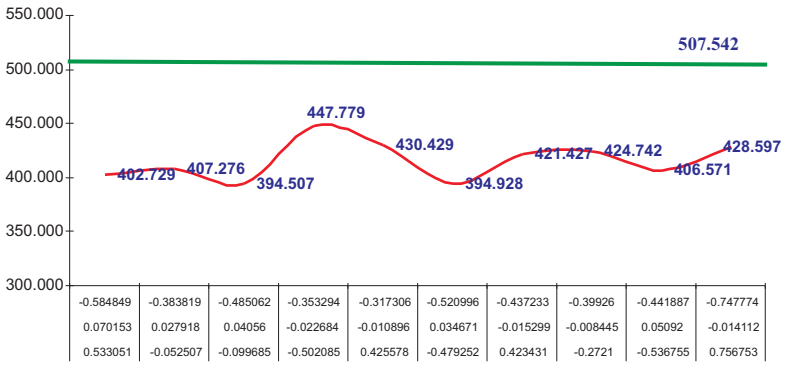


FIGURA 1: Comparación de los resultados obtenidos con el Algoritmo Genético (rojo) y el método clásico (verde), según el AIC.

TABLA 5: Riesgo asociado a cada uno de los parámetros del Modelo de Regresión de Cox

Prueba	$Exp(\beta_1)$	$Exp(\beta_2)$	$Exp(\beta_3)$
1	1,70412367	1,07267229	0,55718999
2	0,94884768	1,02831136	0,68125472
3	0,90512249	1,04139379	0,61565902
4	0,60526736	0,97757135	0,70237067
5	1,53047478	0,98916315	0,72810792
6	0,61924641	1,03527905	0,5939287
7	1,52719237	0,98481744	0,64582094
8	0,76177808	0,99159056	0,67081627
9	0,58464234	1,05223871	0,64282227
10	2,1313445	0,98598711	0,47341921

6. Conclusiones y Recomendaciones

6.1. Conclusiones

- Los A.G. son técnicas de gran utilidad en la optimización de funciones que son difíciles de tratar haciendo uso de las técnicas de optimización clásica.

- El Algoritmo Genético desarrollado permitió encontrar estimaciones de los parámetros del Modelo de Regresión de Cox del Análisis de Supervivencia para pacientes del servicio de Diálisis Peritoneal del Hospital Clínico Universitario de Caracas entre 1980 y 1997, mejores que los encontrados por el método clásico, basándose en la comparación del AIC.

- Un alto porcentaje de los valores de los parámetros del Modelo de Regresión de Cox obtenidos por el Algoritmo Genético, son cercanos a los obtenidos por el método clásico. Puesto que, gran parte de los riesgos asociados a los parámetros caen dentro del intervalo de confianza obtenido pométodo clásico.

6.2. Recomendaciones

- Usar otros criterios para la función de ajuste, especialmente criterios que estudien cada uno de los parámetros por separado, tales como el test de Wald.

- Estudiar de manera más detallada la población inicial, ya que una población inicial heterogénea puede llevar al óptimo global.

- Desarrollar Algoritmos Genéticos para la optimización de los parámetros de otras funciones estadísticas como método alternativo a los métodos clásicos.

Referencias

Borges, R. (2002), Análisis de supervivencia aplicado a un caso de diálisis renal y diálisis peritoneal en el hospital clínico universitario de caracas y hemodiálisis en el hospital de clínicas caracas 1980 - 2000, Tesis de Maestría, Universidad de los Andes, Facultad de Ciencias Económicas y Sociales. Instituto de Estadística Aplicada y Computación. Mérida.

Cox, D. & Oakes, D. (1984), *Analysis of Survival Data*, Chapman and Hall/CRC, New York.