ALGORITMOS GENÉTICOS EN LA DISCRIMINACIÓN

Aurora Montano Rivas[1]
Facultad de Estadística e Informática, Universidad Veracruzana.

ABSTRACT.
The combined use of genetic algorithms with Fisher's Discriminant Function and Logistic Regression is proposed, the goal is to find one or several classification functions. The proposed method was used with a training sample. A value for the discrimination error is fixed as a critical permissible value in an S-Plus coded. It was applied a genetic algorithm to a sample of n models which were generated using resampling method, then yielding new discriminating models. This news models results in less discrimation error than Fisher's Discriminant Function.

RESUMEN.
Se propone el uso combinado de algoritmos genéticos con análisis discriminante de Fisher o análisis de regresión logística con la finalidad de encontrar una o varias funciones de clasificación "óptimas". Con el objetivo de mostrar el funcionamiento de la propuesta, la misma fue utilizada con una muestra de entrenamiento. Se fijó un valor para el error de mala discriminación como valor de tolerancia permisible en un programa escrito en S-Plus. A una muestra de n modelos, generados mediante un mecanismo de remuestreo, se les aplicó un algoritmo genético que produjo nuevos modelos para discriminar. Estos modelos fueron evaluados y mostraron un error de mala clasificación menor que el que mostró el análisis discriminante de Fisher.

KEY WORDS: Genetic Algorithms , Discriminant Analysis, Logistic regression.

MSC: 65C60

1. INTRODUCCIÓN.

La clasificación de objetos o individuos es un problema importante en diversas áreas de la investigación y existe abundante metodología estadística para ello. La clasificación se puede llevar a cabo de dos formas diferentes:

1.- Análisis de clasificación (CLUSTER)

Se agrupan los individuos considerando la descripción que sobre ellos mismos proporcionan ciertas variables independientes.

2.- Análisis discriminante.

Se tiene un grupo de individuos que se sabe pertenecen a ciertas clases o categorías, con la información de ciertas variables se busca una función (función discriminante) que permita "la mejor" separación de esas clases. Un nuevo individuo es asignado a una de las categorías de acuerdo al valor que se obtiene al evaluar la función discriminante en los valores de sus variables.

El análisis discriminante de Fisher hace la suposición de normalidad multivariada y una alternativa, cuando no se cumple ésta, es la regresión logística (Hair, 1999).

La técnica de optimización conocida como "Algoritmos Genéticos" es de utilidad para el fin descrito anteriormente. Propuestos por Holland (1975), son métodos de búsqueda basados en la idea de la evolución natural de los seres vivos. Desde un enfoque genético, se parte de un conjunto de individuos progenitores, los cuales son cruzados y posteriormente mutados, generándose nuevos individuos con aquellas características que más dominaron en los progenitores. Su espacio solución es un conjunto de cromosomas (modelos), los cuales están formados de alelos

[1] julmontano@uv.mx

(*p+1* parámetros). Estos cromosomas son cruzados y los alelos son mutados, con la finalidad de obtener nuevos cromosomas (modelos) con las mejores características.

La aplicación de un algoritmo genético produce una función o un conjunto de funciones que permiten clasificar o discriminar a los individuos con el mínimo error posible. Las funciones (puede ser una sola) así obtenidas se dicen óptimas, o subóptimas.

En este artículo se propone hacer clasificación a través del uso combinado de algoritmos genéticos con análisis discriminante o regresión logística.

Con la finalidad de verificar cuál de estos métodos de optimización (análisis discriminante, y regresión logística combinados con algoritmos genéticos) proporcionan la mejor "función óptima", se realizó un programa en el paquete S-Plus y usando como muestra de entrenamiento la base de datos de Flores de Iris (Fisher, 1936) se propone aplicar análisis discriminante o regresión logística, con algoritmos genéticos con los operadores cruce aritmético y dos mutaciones, una con distribución normal y otra con distribución uniforme. Las muestras para crear los modelos que se necesitan en los algoritmos genéticos, se generan utilizando técnicas de remuestreo. Se presentan las rutinas en S-Plus de los operadores empleados, así como los pasos del programa y los resultados obtenidos.

2. ANÁLISIS DISCRIMINANTE.

El Análisis Discriminante permite asignar o clasificar individuos dentro de g grupos excluyentes o poblaciones previamente definidos. Se basa en el estudio de la relación de una variable dependiente con un conjunto de p variables aleatorias independientes continuas $(X_1, X_2, ...X_p) = X$ las cuales describen a cada uno de los individuos u objetos. Las $g(g-1)/2$ combinaciones lineales o seudolineales de las variables $X_1, X_2, ..., X_p$, permiten la discriminación entre los grupos, de tal forma que la probabilidad de error de mala clasificación sea minimizada o la probabilidad de clasificación correcta sea maximizada (Barbro, Teija y Kaisa, 1996).

Para el caso de g =2, se supone que la población Π_j tiene distribución $N_p(\mu_j, \Sigma)$ para j=1,2 con Σ definida positiva. La combinación lineal es evaluada en los valores de las variables en X , y el resultado se utiliza para definir la regla de decisión en (1). :

$$\begin{aligned} x \in \Pi_1 \quad si \quad a'x > 0 \\ x \in \Pi_2 \quad si \quad a'x \leq 0 \end{aligned} \tag{1}$$

donde $a \in R^{p+1}$ es el vector de coeficientes que minimizan la probabilidad de clasificación errónea definida como:

$$\min\{\frac{1}{2}P(a'x > 0 \mid x \in \Pi_2) + \frac{1}{2}P(a'x \leq 0 \mid x \in \Pi_1)\}.$$

La función así encontrada se llama función discriminante óptima.

3. ALGORITMOS GENÉTICOS.

Los Algoritmos Genéticos son métodos de optimización con búsqueda aleatoria sobre el espacio de posibles soluciones, éstos están inspirados en el proceso evolutivo de Darwin; el cual consiste en seleccionar los mejores cromosomas padres que producirán mejores hijos. El proceso es iterativo, deteniéndose cuando se cumple un criterio de optimalidad, (Michalewicz,1992).
El espacio de posibles soluciones es la población inicial o población cero, la cual está formada por los cromosomas o modelos generados; por la aplicación de alguna de las técnicas como: análisis discriminante, regresión logística,

análisis cluster, regresión múltiple, entre otras. Generalmente el tamaño de esta población es de 20 a 100 cromosomas y se denotaran por n_c, (Banzhaf, Walfgang and Reeves,1999) .

Después de generar la población inicial, cada modelo debe ser evaluado, para ello indicamos los siguientes pasos: (E Mod).

1. La matriz de los datos originales debe quedar particionada en dos grupos de filas tomando como base la variable dependiente. Los datos correspondientes a cada grupo forman una submatriz.

2. Se debe conocer el tamaño de Π_1 y Π_2

3. Cada modelo i será evaluado con los datos originales (i=1,2,… ,n_c).

4. Obtener la proporción π_1 de valores negativos dentro de Π_1 y la proporción de valores positivos π_2 dentro de Π_2.

5. Calcular

$$\frac{\sum_{i=1}^{2} \pi_i}{2} = e_i ,$$

el cual se llamará error de discriminación promedio del modelo i.

6. Incrementar i y regresar al paso 3.

De esta manera se obtienen los errores promedios de cada función objetivo, para proceder a seleccionar los modelos.

3.1 Método de Selección de los modelos.

Existen muchos métodos para la selección de modelos progenitores, en este caso se consideró el método de la ruleta, el cual los elegirá de acuerdo a una probabilidad basada en la función objetivo, es decir, los individuos con mejor valor ajustado tienen mayor oportunidad de ser seleccionados.

Pasos a seguir para seleccionar los modelos:

1. Calcular $T = \sum_{i=1}^{n_c} e_i$.

2. Obtener el fitness o probabilidad de ajuste para cada modelo: $p_i = \frac{e_i}{T}$.

3. Obtener para cada modelo, la probabilidad de ajuste acumulada: $c_i = \sum_{j=1}^{i} p_j$.

4. Generar n_c valores aleatorios $U(0,1)$ y ordenarlos de manera ascendente: $u_{(1)}, u_{(2)}, ..., u_{(n_c)}$, tal $u_{(i)} \leq, u_{(j)}$ para $i < j$.

5. El modelo i será seleccionado y copiado a la nueva población si cumple la condición (1):

$$C_{i-1} < u_{(i)} < C_i \qquad i=1,2,\ldots,n_c \qquad (2)$$

Con $C_0=0$

3.2 Operadores Genéticos para Datos Continuos.

Los operadores genéticos proporcionan los mecanismos de búsqueda de los algoritmos genéticos, son usados para crear nuevas soluciones, basados en el conjunto de los mejores modelos elegidos previamente (Houck, Joines y Kay1997). Se clasifican en:

Cruce aritmético:

En este caso dos modelos o cromosomas son cruzados para generar dos nuevos. Esto se hace con la siguiente rutina en SPlus.

Sea M la matriz correspondiente a los modelos previamente seleccionados, los cuales serán cruzados. Identifíquese por i, j los modelos que serán cruzados y M[i,] la fila i de dicha matriz. Los índices i=1, 2,. . .,h identifican a los diferentes modelos.

Rutina Cruce Aritmético (RCA).

```
k <- 0
for(i in 1:h){
    for(j in 1:h){
        if(i<j){
            r <- runif(1,min=0,max=1)
            vy <- r*M[i,]+(1-r)*M[j,]
            k <- k+1
            MCOE[k,]<- vy
            vx <-(1-r)*M[i,]+r*M[j,]
            k <- k+1
            MCOE[k,]<- vx
        }
    }
}
```

Como resultado de un cruce aritmético se obtiene una matriz de nuevos modelos.

Mutación:

Se realiza después de un cruce, con la finalidad de prevenir posibles fallas de las soluciones en la población. Con este operador un gene (vector columna de la matriz M) es seleccionado aleatoriamente, y sufre una alteración y se produce una nueva solución que es única.

En este caso se presentan dos tipos de mutaciones:

a) Rutina elaborada en Splus para la mutación con valor aleatorio de una Distribución Normal: (RMDN).

```
repeat{
    p1 <- p+1
    z <- sample(1:p1,1)
    xbar <- mean(MCOE[,z])
    desv <-stdev(MCOE[,z])
```

```
    if(desv != 0)
        break
  }
w <- rnorm(1, xbar, desv)
MCOE[,z]<-c(rep(w,k))
```

b) Runita elaborada en Splus para la mutación con valor aleatorio de una Distribución Uniforme: (RMDU)

```
repeat{
    p1 <- p+1
    z <- sample(1:p1,1)
    if(min(MCOE[,z])!= max(MCOE[,z])0)
        break
    }
w <- runif(1, min(MCOE[,z]), max(MCOE[,z]))
MCOE[,z]<-c(rep(w,k))
```

4. DESCRIPCIÓN DEL PROCEDIMIENTO.

En el problema de la discriminación para el caso de $g = 2$, se sabe a cual población pertenecen los individuos, y las variables independientes deben cumplir el supuesto de normalidad y homogeneidad de varianzas, mientras que en el problema de clasificación a nuevos individuos se les miden las mismas características que a los anteriores, y se quiere conocer a cuál población pertenecen. El problema, en su totalidad, se resuelve usando Análisis Discriminante, en caso de que los supuestos requeridos no se cumplan, se usa Regresión Logística.

Estas dos metodologías permiten "minimizar" el error, pero ¿Existirá una técnica que permita "minimizar" más ese error de clasificación?

La propuesta es usar Algoritmos Genéticos, porque permite realizar búsquedas que pueden resultar bastante "exhaustivas". Es decir, permite explorar un conjunto de funciones discriminantes o logísticas generadas, para determinar cuál o cuáles de ellas son las que presentan el menor error de clasificación, en comparación con las técnicas tradicionales. Para ello, se propone hacer remuestreo de la muestra de entrenamiento, y a cada nueva muestra se le aplica análisis discriminante o regresión logística, para obtener los modelos que forman la población inicial. También se propone usar la mutación con un valor aleatorio de una distribución normal, para ver si presentan mejores resultados que los proporcionados por mutación uniforme; además se realiza la validación usando el método *Estimación a partir de datos propuestos* (Dallas, 1998).

5. UN CASO DE ESTUDIO.

Se realizó un programa en el paquete Estadístico S-Plus (versión 6.1) y se usó como muestra de entrenamiento la base de datos de dos especies (las que presentan interacción) de Flores de Iris (Fisher, 1936); la cual está constituida por 100 individuos, 50 de cada especie y 4 variables independientes:

Y	=	Especies (Versicolor y Virginica).
X_1	=	Longitud del sépalo
X_2	=	Ancho del sépalo
X_3	=	Longitud del pétalo
X_4	=	Ancho del pétalo

Búsqueda del óptimo.

1) La muestra de entrenamiento se captura en el paquete Estadístico S-Plus, donde la variable dependiente, debe ser declarada de tipo factor (variable categórica en el S-Plus).

2) Aplicar un análisis discriminante o una regresión logística, y el error promedio de discriminación se usará en el programa como el valor de tolerancia.

3) La base de datos se divide en dos muestras A y B ; para el ejemplo fueron cada una de tamaño 50, de manera que cada una de éstas contiene 25 individuos de cada especie.

4) Se usa la muestra A para generar los modelos, la cual para este caso de validación pasa a ser la matriz de datos originales.

5) Se aplicar remuestreo y con cada muestra resultante se hace un análisis discriminante o regresión logística

6) Se evalúa cada modelo, usando la muestra A. (Según E Mod)

7) Se aplica el método de selección, descrito en 3.1.

8) Se hace el cruce aritmético. (utilizando RCA)

9) Se evalúa cada modelo, usando la muestra A. (Según E Mod)

10) Se evalúan los modelos usando la muestra B. (Según E Mod)

11) Se aplica la mutación normal. (RMDM)

12) Se evalúa cada modelo usando la muestra A. (Según E Mod)

13) Se evalúan los modelos usando la muestra B. (Según E Mod)

14) Se aplica la mutación uniforme. (RMDU)

15) Se evalúa cada modelo usando la muestra A. (Según E Mod)

16) Se evalúan los modelos usando la muestra B. (Según E Mod)

Si en esta etapa no se encuentra el óptimo o subóptimo, se regresa al paso 7. El programa termina cuando encuentra modelos con error promedio menor que la tolerancia dada.

En los pasos 9, 12 y 15, se hacen evaluaciones para ver los errores de discriminación y en los pasos 10, 13 y 16 se observan los errores de clasificación. Esto con la finalidad de determinar en qué momento son mejores y cuándo se logra obtener la función óptima.

6. RESULTADOS.

Al ejecutar el programa sin validación se obtuvo el análisis discriminante de Fisher para la base completa, obteniéndose el modelo original (Cuadro 1 del ANEXO) con error promedio de 0.03, el cual se tomó como tolerancia en el programa. Se aplicó el remuestreo a la muestra de entrenamiento, obteniéndose 20 muestras aleatorias y a cada una se le aplicó el análisis discriminante; por lo que la población inicial se formó de 20 modelos con 5 variables.

Se aplicó la técnica de algoritmos genéticos, obteniendo dos modelos por mutación normal y otros dos por mutación uniforme, ambos operadores proporcionaron funciones con error promedio de 0.02, menor que el proporcionado por Fisher. (ver Cuadro 2 y 3)

De la misma manera se procedió usando regresión logística y se encontró que el error promedio es muy alto (ver Cuadro 4), por lo que difícilmente un modelo de regresión logística logrará cumplir con la tolerancia fijada por

discriminación de Fisher, (0.03) y mucho menos reporte una función con error igual a 0.02; de manera que se decidió probar con una tolerancia de e=0.70 generando n=20 remuestreos.

Aplicando la técnica propuesta se tiene que por mutación normal 19 funciones reportan un error menor que la tolerancia, en 15 de ellas disminuyó hasta 0.50. Para el caso de mutación uniforme todos los modelos reportan un error de 0.50, lo que implica que sólo un 50% de los individuos logran ser clasificados adecuadamente.

Como se mencionó anteriormente, la base de datos original se dividió en dos partes, muestra 1 y muestra 2; el análisis discriminante de la muestra 1 reportó un error promedio de 0.02, (ver Cuadro 5) por lo que se espera encontrar una función con error promedio de discriminación menor o igual que éste.

Aplicando el remuestreo se encontró un modelo con error promedio de cero, es decir, se encontró la función óptima, lo que nos llevaría a detener el proceso, pero como el objetivo es validar la técnica propuesta, se continúa la ejecución del programa.

La mutación normal proporcionó dos modelos con errores promedio de discriminación de 0.00 y 0.02., y por mutación uniforme fueron 0.02 y 0.00; el error promedio es el mismo pero los coeficientes de los modelos son muy diferentes. Dichos modelos se evaluaron con los datos de la muestra 2, y se encontró que los errores de los modelos por mutación normal aumentaron hasta 0.08 y 0.04 respectivamente; mientras que los obtenidos por mutación uniforme presentan un error de clasificación similar (0.0016).

Por otra parte, la validación para regresión logística no funcionó, ya que los errores de clasificación fueron muy grandes.

7. CONCLUSIONES.

Los métodos de clasificación comúnmente utilizados son el análisis discriminante y el de regresión logística, y se recomienda emplearlos cuando cumplen o no respectivamente el supuesto de normalidad. En las últimas décadas se aplican los métodos de algoritmos genéticos y al combinarlos con el análisis discriminante o con regresión logística se obtienen modelos de clasificación más eficientes. Para lo anterior se realizó el programa Dislog en el paquete estadístico Splus, en el cual se ejecutó una base de datos de Flores de Iris con 100 observaciones, cuatro variables independientes y una dependiente o de agrupación. Se obtuvieron los modelos de clasificación y sus respectivos errores de tolerancia y en relación a ella se concluye que la técnica de algoritmos genéticos aplicada a las funciones proporcionadas por el análisis discriminante, genera los mejores modelos discriminantes con el mínimo error de clasificación; es decir, los modelos obtenidos son más eficientes que los reportados por el método de Fisher.

También se encontró que al realizar la validación de la técnica, los modelos proporcionados por la mutación normal, son muy buenos discriminando, pero no tanto al realizar una clasificación, mientras que los presentados por mutación uniforme no reportan cambios significativos al pasar de la discriminación a la clasificación.

Received April, 2007
Revised October 2007

REFERENCIAS

[1]BANZHAF, WALFGANG and REEVES, COLIN (1999). **Foundations of genetic algorithms**, Morgan Kaufmann Publishers, Inc, San Francisco, California.

[2]BARBRO; B., TEIJA, L. Y KAISA, S. (1996). Choosing Bankruptcy Predictors Using Discriminant Analysis, Logit Analysis, and Genetic Algorithms. **Centre for Computer Science, Finland. Technical. Report No. 40.**

[3] DALLAS, E. J. (1998). **Métodos multivariados aplicados a los análisis de datos**. Internacional Thomson Editores, Madrid, España.

[4] FISHER R. A. (1936). The Use of Multiple Measurements in Taxonomic Problems. **Annual Eugenics.** 7:179-188.

[6] HOUCK, C. R. JOINES, J. A., y KAY, M. G. 1997. **A Genetic Algorithm for Function Optimization: A Matlab Implementation.** North Carolina State University.

[7] MICHALEWICZ, Z.(1992). **Genetic Algorithms + Data Structures =Evolution Prog**rams. Springer. New York.

ANEXO.

Cuadro 1. Función discriminante óptima por el método de Fisher.

Longitud Sépalo	Ancho Sépalo	Longitud Pétalo	Ancho Pétalo	Término. Const.	Error Grup1	Grup2	Promedio.
3.556303	5.578621	-6.970128	-12.38604	16.66309	0.04	0.02	0.03

Cuadro 2. Funciones generadas por la Mutación (Distribución Normal).

Longitud Sépalo	Ancho Sépalo	Longitud Pétalo	Ancho Pétalo	Término Constante	Error Promedio
4.389463	4.311353	-7.96471	-12.08190	18.80822	0.02
3.937811	7.306391	-7.96471	-13.45358	16.23628	0.02

Cuadro 3. Funciones generadas por la Mutación (Distribución Uniforme).

Longitud Sépalo	Ancho Sépalo	Longitud Pétalo	Ancho Pétalo	Término Constante	Error Promedio
4.367464	6.900467	-8.98465	-14.31159	21.73694	0.02
4.367464	5.653616	-8.68388	-11.99341	18.91315	0.02

Cuadro 4. Función de regresión logística de la base de datos original.

Longitud Sépalo	Ancho Sépalo	Longitud Pétalo	Ancho Pétalo	Término Const.	Error Grup1	Error Grup2	Error Promedio
-2.4652	-6.68065	9.42901	18.2855	-42.63567	0.98	0.98	0.98

Cuadro 5. Función discriminante de Fisher de la muestra 1.

Longitud Sépalo	Ancho Sépalo	Longitud Pétalo	Ancho Pétalo	Término Const.	Error Grup1	Error Grup2	Error Promedio
3.91284	5.190329	-7.63229	-12.7248	19.5235	0.04	0	0.02

Cuadro 6. Solución inicial con error cero.

Longitud Sépalo	Ancho Sépalo	Longitud Pétalo	Ancho Pétalo	Término Const.	Error Promedio
2.556413	3.513466	-7.55458	-11.84034	31.50987	0.00

Cuadro 7. Solución por Mutación por distribución normal.

Long Sépalo	Ancho Sépalo	Long Pétalo	Ancho Pétalo	Térm Const.	Error Promedio
2.7516019	4.954462	-7.793511	-14.16819	31.99538	0.00
4.8587625	4.954462	-9.235785	-13.31772	22.73018	0.02

Cuadro 8. Solución por Mutación por distribución uniforme.

Long Sépalo	Ancho Sépalo	Long Pétalo	Ancho Pétalo	Térm Const.	Error Promedio
5.324897	5.38691	-9.576163	-13.61683	20.95264	0.02
4.858762	5.38691	-9.235785	-13.31772	22.73018	0.00