Algoritmos Genéticos

Gonzalo V. Castiglione, Alan E. Karpovsky, Martín Sturla
Estudiantes Instituto Tecnológico de Buenos Aires (ITBA)

12 de Junio de 2012

Entrega Final - Informe

***Resumen*—El presente informe busca analizar la utilización de algoritmos genéticos para la obtención de pesos óptimos para redes neuronales multicapas que aproximen funciones reales en intervalos acotados. Se estudiarán distintas técnicas de selección, cruza, mutación y reemplazo de los individuos y se detallarán los resultados obtenidos.**

***Palabras clave*—algotitmos genéticos, métodos de selección, métodos de cruza, redes neuronales, evolución, población, mutación, reemplazo, individuo, cromosoma**

I. Introducción

Se analizó el comportamiento de los algoritmos genéticos en el problema de la obtención de pesos para redes neuronales multicapa que aproximan funciones reales.
Con este fin se implementó un algoritmo que permite definir de manera sencilla todos los parámetros más importantes del motor de algoritmos genéticos con el fin de hacer más simple y práctico su estudio. Los parámetros que el usuario puede modificar son los siguientes:

- Tamaño de la población (N)
- Brecha generacional (G)
- Número máximo de generaciones
- Probabilidad de mutación (p_m)
- Probabilidad de cruce (p_c)
- Método de selección
- Método de reemplazo

II. Desarrollo

A. Modelado del problema

A.1 Representación de los individuos

Se decidió representar a cada individuo de la población (red neuronal multicapa en este caso) de la siguiente forma: Cada capa de la red neuronal está represnetada por una matriz de pesos; por consiguiente una red neuronal es la concatenación de los las filas de la matriz de cada una de sus capas. Un cromosoma será entonces un arreglo de números reales en el que cada locus representará un peso puntual de la red (notar que el *bias* esta representado como una conección extra a cada una de las neuronas).

En otras palabras, cada *locus* representa una conexión en la red neuronal. Su *alelo* respectivo, un número real, indica el peso de dicha conexión. Para reconstruir la red a partir del cromosoma se guarda convenientemente la arquitectura utilizada en otro lado.

A.2 Diagramación del algoritmo

En lo que a algoritmos genéticos respecta, existen numerosas formas de pensar la diagramación del mismo. Si bien la base teórica de fondo siempre es la misma, la forma en la que se eligen los individuos para ser cruzados, la forma en la que los individuos cruzados pasan o no a la siguiente generación y demás aspectos pueden quedar a elección de quien sea que implemente el motor de algoritmos genéticos.

En la **Figura 1** del **Anexo A** se observa el modelado elegido para la implementación del algoritmo y en la **Figura 2** se ejemplifica un caso particular:

Suponiendo poblaciones de 10 individuos y un gap generacional $G = 0{,}6$ se seleccionan, mediante alguno de los métodos de selección que se desarrollarán a continuación, un $(G * 100)\,\%$ de la cantidad de individuos (N) de la *generación i*. Es decir que para el ejemplo de poblaciones de 10 individuso y $G = 0{,}6$, se tomarán 6 individuos (un 60 % del tamaño de la población). Estos 6 individuos seleccionados son cruzados mediante alguno de los métodos de *crossover* y luego son pasados directamente a la *generación i+1* con cierta probabilidad de mutación y/o de *backpropagation*. Esto quiere decir que con cierta probabilidad baja, los hijos de los individuos seleccionados pueden ser mutados y/o refinados mediante *backpropagation* antes de ser pasados a la próxima generación.

De esta forma, la *generación i+1* ya tiene 6 de los 10 individuos necesarios. Los 4 individuos restantes son seleccionados de entre los 10 individuos de la *generación i*; en otras palabras son seleccionados entre los padres de los hijos obtenidos por la cruza y los que nunca fueron elegidos en un principio por el método de selección. Nótese que los 4 individuos en cuestión del ejemplo son seleccionados mediante los métodos de reemplazo.

El motor de algoritmos genéticos implementado tiene la particularidad de permitir el uso de métodos mixtos para la selección y el reemplazo de los individuos. El usuario puede

12 de Junio, 2012.

optar por elegir los individuos a cruzar bajo el método de selección *Elite* y luego elegir los individuos a reemplazar utilizando *Ruleta*.

B. Función de fitness

La función de *fitness* mide el grado de adaptación de un determinado individuo al entorno actual. Para este problema en particular se optó por tomar como función de *fitness* $f(i) = \frac{1}{ECM}$ siendo i una red neuronal (individuo) y ECM el error cuadrático medio obtenido al evaluar la misma.

III. Métodos de selección y reemplazo

Los métodos de selección y reemplazo son utilizados, valga la redundancia, para seleccionar los individuos de una determinada población. El *input* de este tipo de métodos es básicamente una población y algún parámetro de configuración (como ser la brecha generacional) y el *output* de éstos es un conjunto determinado de individuos.

A. Elite

El método de selección/reemplazo *Elite* consiste en elegir a los k "mejores" individuos de la población, considerando mejores a los individuos con mejor *fitness*.

El método de selección elite es algo distinto a los demás. Se eligen los k individuos a cruzar. Se cruzan para obtener unos k nuevos individuos. De esos $2k$ individuos, se eligen los mejores k para la nueva generación. En otras palabras, si los k hijos producidos tienen menor fitness que sus padres, ninguno de ellos pasa a la nueva generación.

B. Ruleta

La selección/reemplazo por ruleta se realiza de la siguiente manera:

1. Se evalúa el fitness, fi, de cada individuo de la población.
2. Se computa la probabilidad *(slot size)*, p_i, de seleccionar al miembio i de la población: $p_i = \frac{f_i}{\sum_{j=1}^{N} f_j}$, donde N es el tamaño de la población.
3. Calcular la probabilidad acumulada, q_i, para cada individuo: $q_i = \sum_{j=1}^{i} p_j$.
4. Generar un número *random*, $r \in (0, 1]$.
5. Si $r < q_1$ entonces seleccionar al primer cromosoma, x_1. Sino seleccionar al individuo x_i tal que $q_{i-1} < r \leq q_i$.
6. Repetir los pasos 4 y 5 k veces para crear k candidatos seleccionados.

C. Boltzman

La selección/reemplazo por Boltzman estipula que la probabilidad de ser elegido es proporcional a una función no lineal del *fitness* y de la "temperatura". En este sentido guarda ciertas similitudes a *simmulated annealing*: al principio busca diversidad (la probabilidad de elegir cada individuo es más uniforme) y luego baja la temperatura haciendo que cada vez haya menos diversidad de acuerdo a una cierta función decreciente monótona de la temperatura.

La probabilidad de cada individuo i de ser elegido es de:

$$\frac{e^{\frac{f_i}{T}}}{\sum_i e^{\frac{f_i}{T}}}$$

Donde T es la temperatura, f_i el fitness de cada individuo. Es fácil ver que si T tiende a infinito, todos los términos tienden a 1, por lo que la probabilidad de cada individuo de ser elegido es $\frac{1}{N}$, es decir uniforme. Por otro lado, a medida que T tiende a valores más pequeños, el peso de los términos con mayores valores de f_i crece exponencialmente en relación con el resto, por lo que la probabilidad de ser elegidos sube.

Se debe tener cuidado con la elección de la temperatura T en relación al rango de posibles valores de fitness. Esto se debe a que si la razón $\frac{f_i}{T}$ toma valores grandes, puede haber problemas de representación y almacenamiento de el número e elevado a dicho valor. En particular se decidió utilizar una temperatura mínima de 1000, dado que asumiendo un fitness de 10000 correspondiente a un error cuadrático medio de 10^{-5} el coeficiente vale 10 y no trae los problemas ya mencionados.

D. Torneo

En la selección/reemplazo por torneo se procede de la siguiente forma:

1. Se eligen 2 individuos al azar.
2. Se toma un número *random* $r \in [0, 1]$.
3. Si $r < 0{,}75$ se seleciona al más apto (de mayor *fitness*), sino se selecciona al menos apto.
4. Ambos individuos se devuelven a la población original y podrían ser seleccionados nuevamente.

E. Universal

La selección universal estocástica se asimila mucho a ruleta pero a diferencia de ésta se genera un solo $r \in [0, \frac{F}{k}]$ para elegir k individuos. Se tiene a su vez que $r_j = \frac{r+j-1}{k}$ con $j \in [1, k]$.

F. Mixto

La selección/reemplazo mixto consiste en elegir k_e individuos utilizando *Elite* (ke es ingresado como parámetro) y el resto de los individuos por *Ruleta / Boltzman*.

IV. Criterios de corte

Los criterios de corte implementados son los siguientes:

- **Máxima cantidad de generaciones:** Dado un número p, el algoritmo termina al alcanzarse p generaciones.
- **Entorno al óptimo:** Se llega a la solución óptima o se alcanza un fitness superior a una determinada cota.
- **Contenido:** Se corta al detectar que el mejor fitness de la población no progresa con las generaciones.
- **Estructura:** Se finaliza al detectar que una parte relevante de la poblacioón no cambia de generacioón en generación. Es decir, dado un porcentaje p, el algoritmo termina cuando la cantidad de individuos iguales de la generación es mayor a dicho p.

V. Mutación y refinamiento

Una vez cruzados, los individuos pasan a la siguiente generación con una probabilidades bajas e independientes de ser mutados y/o refinados mediante el algoritmo de *backpropagation*.

A. Mutación

Es importante entender que existe una probabilidad de que se produzca la mutación de un individuo y en caso de que ésto sea verdadero se puede producir la mutación de cada locus de ese individuo. Sin embargo, si la cantidad de locus es considerable, este procedimiento puede ser innecesariamente ineficiente, teniendo que generar $L+1$ números al azar, siendo L la cantidad de locus, si sucede el primer evento. Sin embargo, si se considera p_m la probabilidad de que mute un individuo, y p la probabilidad de que mute cada locus, se verifica que la probabilidad de que no mute ningún locus es:

$$P(X=0) = p_m(1-p)^L + (1-p_m)$$

En otras palabras, equivale a la probabilidad de que el individuo no mute más la probabilidad de que mute pero no mute ningún locus. La notación variable probabilística X denota la cantidad de locus mutados. También se puede verificar que la probabilidad de que muten exactamente i locus está dada por:

$$P(X=i) = p_m p^i (1-p)^{L-i} comb(L,i)$$

Sin embargo, si se toma p_m y p pequeños, el factor $p_m p^i$ se hace arbitrariamente pequeño, por lo que las probabilidades disminuyen drásticamente para valores incrementales de i. Se podría incluso despreciar la probabilidad de que muten dos locus o más. Dicha probabildid está dada por:

$$P(X \geq 2) = 1 - (P(X=0) + P(X=1))$$

$$P(X \geq 2) = 1 - (p_m(1-p)^L + (1-p_m) + p_m p(1-p)^{L-1} L)$$

$$P(X \geq 2) = -p_m(1-p)^L + p_m - p_m p(1-p)^{L-1} L)$$

$$P(X \geq 2) = p_m \left(1 - (1-p)^L - p(1-p)^{L-1} L)\right)$$

Por ejemplo, para $L = 50$, $p_m = 0{,}01$ y $p = 0{,}001$, esta expresión es aproximadamente 10^{-5}, mientras que la probabilidad de que mute sólo un locus es $5\text{x}10^{-3}$. En vista de estos resultados, se puede despreciar la probabilidad de que mute más de un locus y definir una nueva probabilidad p', tal que p' representa la probabilidad de que mute algún locus, y $1-p'$ es la probabilidad de que no haya mutación.

Dadas las consideraciones anteriores, la mutación se realiza de la siguiente manera: se genera un número al azar entre 0 y 1. Si es menor a p', se toma un locus al azar del individuo y se lo modifica. A pesar de que estrictamente no es equivalente al modelo de mutación ya explicado, es más eficiente.

La modificación del locus consiste en generar un un número al azar proporcional al valor del mismo y otro random no proporcional al valor. La idea detrás de esto es que el ruido pueda afectar de igual a igual a valores grandes como muy pequeños. En conclusión, la mutción consiste en agregar cierto ruido a uno de los pesos de la red neuronal en cuestión. Cabe destacar que los cálculos asumen que la cantidad de locus (L) no es demasiado grande, caso contrario la probabilidad de que mute más de un locus deja de ser despreciable (la distribución binomial se empieza a comportar como una distribución normal). En otras palabras, si la arquitectura es particularmente grande se debería reconsiderar dicho método.

A.1 Mutación clásica

Consiste en lo ya explicado en la sección anterior. Si se obtiene un número al azar entre 0 y 1 menor a p' se elige un locus al azar para mutar.

A.2 Mutación no uniforme

Similar a la mutación clásica, con la salvedad que p' no es constante sino que pasa a ser una función de las generaciones elapsadas. En otras palabras se tiene una función $p'(g)$ decreciente. La idea de que dicha función sea decreciente es similar al objetivo de la selección de *Boltzman*: aumentar la diversidad en generaciones tempranas y reducirla en generaciones más tardías. En particular se utilizó la siguiente función:

$$p'(g) = c^g p'_0$$

Donde c es algún coeficiente real positivo menor que 1.

B. Refinamiento

Una vez finalizada la etapa de mutación, se genera un nuevo número al azar entre 0 y 1. Si dicho número es menor a una probabilidad p_b, se refina el individuo generado. La etapa de refinamiento consiste en un algoritmo de backpropagation con una cantidad de etapas acotadas.

VI. Resultados

Si se observa con detenimiento la **Tabla 1** del **Anexo B**, se puede observar los resultados obtenidos por el algoritmo genético con distintos criterios de selección y reemplazo, utilizando un criterio de corte de 100 generaciones. Se puede apreciar que todos los criterios fueron exitosos en minimizar hasta cierto extento el error cuadrático medio de la red, alcanzando al menos un error en el orden de 10^{-3} en el mejor individuo. De los resultados se pueden apreciar que los métodos *Elite* son los más rápidos en converger a menores valores, sin embargo pagan el precio de tener una menor diversidad. Esto es evidente si se compara el error de generalización del mejor individuo con el error de generalización promedio de los individuos, los cuales son particularmente similares (es una condición necesaria pero no suficiente, se podrían tener individuos tan similares pero pesos muy distintos, aunque es poco probable).

Esto causa que los métodos que hacen mucho uso de elitismo se estanquen en mínimos locales fácilmente. Por ejemplo, si se hace observa en la misma tabla el resultado del método de selección *Elite* con reemplazo *Mixto Elite-Boltzman*, se alcanzaron valores de $4{,}2\text{x}10^{-4}$, $4{,}5\text{x}10^{-4}$ y

$4{,}2\text{x}10^{-4}$ para el error cuadrático medio de entrenamiento del mejor individuo, el de generalización del mejor individuo, y el de generalización promedio de todos los individuos respectivamente. Se decidió entrenar una red con la misma configuración por 5000 generaciones, y el error mejoró a tan solo $9{,}1\text{x}10^{-5}$, $3{,}6\text{x}10^{-4}$ y $3{,}6\text{x}10^{-4}$ respectivamente. Esto deja en evidencia que dudosamente se pueda obtener un error sustancialmente menor al obtenido en unas 100 generaciones utilizando métodos muy elitistas.

Por otro lado, los métodos que no son tan elitistas no alcanzan en pocas generaciones errores tan bajos como los métodos más elitistas. Sin embargo sí hacen que la diversidad sea mayor. Nuevamente eso se puede apreicar en la diferencias entre el error alcanzado por el mejor individuo y el error promedio de los individuos. Si se observa el caso del método de Boltzman para tanto la selección como el reemplazo, se puede apreciar que el error del mejor individuo está en el orden de 10^{-3} mientras el error promedio es de 0,011. Esto indica que métodos que promueven la diversidad como lo es el de Boltzman en generaciones tempranas hará que se tenga una diversidad muy grande de individuos, lo cual muy probablemente se traduzca a tener valores muy variados de fitness en la población. Por lo general este tipo de métodos también suelen alcanzar errores en el orden de 10^{-4} pero no antes de alcanzar al menos 1000 generaciones.

Sin importar qué método se haya utilizado, no se encontraron casos que reduzcan el error a órdenes inferiores a 10^-4, lo cual podría indicar quizás una cota para la arquitectura analizada.

Por otro lado, se observó que el refinamiento mejoró sustancialmente la efectividad del algoritmo. Sin embargo, el refinamiento conlleva a que el tiempo de ejecución de las generaciones sea mayor. Debido a esto, si se decide no utilizarlo se puede, por ejemplo, utilizar una población mucho mayor sin una pérdida considerable de performance.

VII. Conclusión

Tras el análisis de los resultados se pueden formular a las siguientes conclusiones:

- No existe un método de selección o reemplazo óptimo. Todos conllevan sus respectivas ventajas o desventajas, y la efectividad de cada uno depende de la naturaleza del problema estudiado.
- Utilizar métodos de selección y reemplazo muy elitistas acelera la convergencia inicial a muestras con mayor fitness, pero reduce particularmente la diversidad de la población y por lo tanto puede caer en mínimos locales. Se consideró que debido a esta limitación el criterio de corte por *Contenido* resultó muy adecuado para estos casos.
- Los criterios que utilizan cierto elitismo junto con algún método que aumente la diversidad de la población traen buenos resultados a largo plazo. Se observó además que dichos métodos son también consistentes a largo plazo, es decir que en sucesivas ejecuciones los resultados son particularmente similares. Esto deja en evidencia, a diferencia de los métodos elititistas que caen fácilmente en mínimos locales, que dicha combinación de métodos conlleva un mejor cubrimiento del dominio.
- Si existe la posibilidad de navegar la función de costo utilizando el gradiente, como es el caso de utilizar el algoritmo de *backpropagation*, se puede implementar un refinamiento de la población que mejora sustancialmente el fitness de la población.

Anexo A: Gráficos

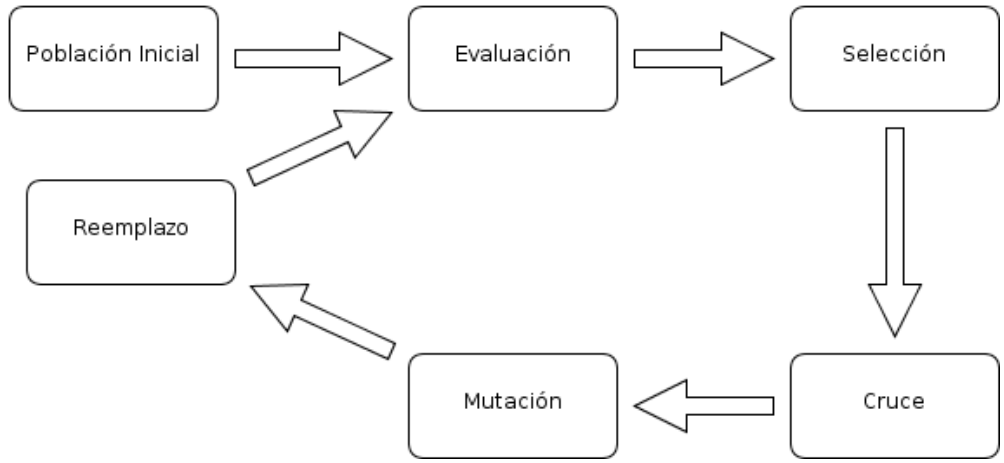

Figura 1: Flujo y etapas del algoritmo genético implementado.

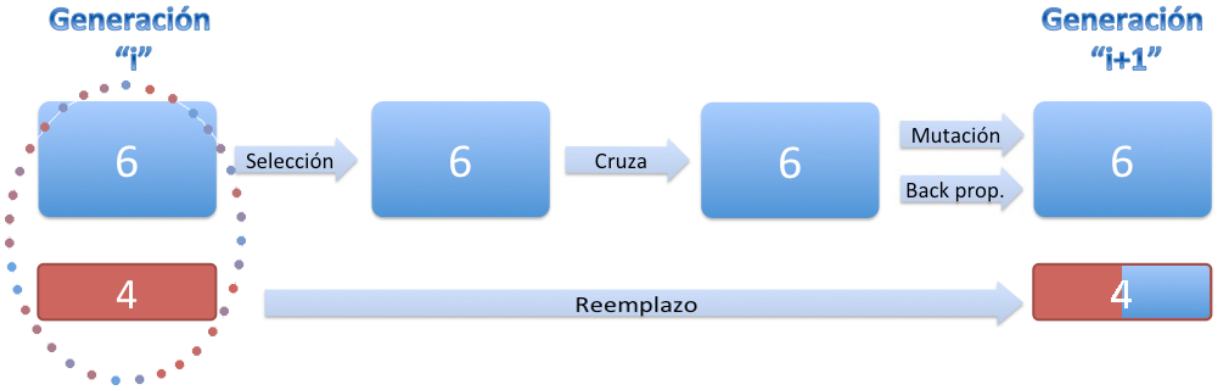

Figura 2: Modelado esquemático del funcionamiento del algoritmo genético para poblaciones de 10 individuos y un gap generacional G = 0.6.

Anexo B: Tabla de Resultados

Selección	Reemplazo	Min. Entrenamiento	Min. Generalizacion	Prom. Generalización
Ruleta	Elite	$1{,}6\text{x}10^{-3}$	$2{,}5\text{x}10^{-3}$	$3{,}0\text{x}10^{-3}$
Boltzman	Elite	$6{,}4\text{x}10^{-4}$	$8{,}9\text{x}10^{-4}$	$1{,}3\text{x}10^{-3}$
Universal	Elite	$1{,}7\text{x}10^{-3}$	$1{,}9\text{x}10^{-3}$	$2{,}1\text{x}10^{-3}$
Torneo	Elite	$7{,}8\text{x}10^{-4}$	$9{,}7\text{x}10^{-4}$	$1{,}1\text{x}10^{-3}$
Elite	Elite	$2{,}3\text{x}10^{-4}$	$3{,}2\text{x}10^{-4}$	$3{,}2\text{x}10^{-4}$
Ruleta	Mixed-Boltzman	$3{,}2\text{x}10^{-3}$	$4{,}3\text{x}10^{-3}$	$6{,}0\text{x}10^{-3}$
Boltzman	Mixed-Boltzman	$1{,}1\text{x}10^{-3}$	$8{,}1\text{x}10^{-4}$	$1{,}4\text{x}10^{-3}$
Universal	Mixed-Boltzman	$8{,}3\text{x}10^{-4}$	$8{,}3\text{x}10^{-4}$	$8{,}3\text{x}10^{-4}$
Elite	Mixed-Boltzman	$4{,}2\text{x}10^{-4}$	$4{,}5\text{x}10^{-4}$	$4{,}2\text{x}10^{-4}$
Ruleta	Boltzman	$1{,}7\text{x}10^{-3}$	$2{,}4\text{x}10^{-3}$	$3{,}7\text{x}10^{-3}$
Boltzman	Boltzman	$6{,}9\text{x}10^{-3}$	$7{,}3\text{x}10^{-3}$	0,011
Elite	Boltzman	$2{,}8\text{x}10^{-4}$	$6{,}3\text{x}10^{-4}$	$6{,}3\text{x}10^{-4}$
Universal	Boltzman	$1{,}8\text{x}10^{-3}$	$2{,}3\text{x}10^{-3}$	$2{,}4\text{x}10^{-3}$
Ruleta	Ruleta	$1{,}4\text{x}10^{-3}$	$1{,}7\text{x}10^{-3}$	$1{,}7\text{x}10^{-3}$
Elite	Ruleta	$8{,}9\text{x}10^{-4}$	$9{,}5\text{x}10^{-4}$	$9{,}5\text{x}10^{-4}$
Boltzman	Ruleta	$1{,}2\text{x}10^{-3}$	$2{,}2\text{x}10^{-4}$	$2{,}4\text{x}10^{-4}$
Universal	Ruleta	$5{,}2\text{x}10^{-3}$	$4{,}7\text{x}10^{-3}$	$5{,}2\text{x}10^{-3}$

Tabla 1: Datos de ejecución para los distintos criterios de selección.

- 30 cromosomas
- Brecha generacional 0,5.
- 100 Generaciones (criterio de corte)
- Probabilidad de mutación (clásica): 0,02.
- Probabilidad de cruce: 0,4. Crossover anular.
- Probabilidad de refinamiento: 0,05.
- $k_e = 6$ (para los ejemplos de *Mixed*).
- Red con dos capas ocultas, ambas de 15 neuronas.
- Función de activación tangente hiperbólica.