Benemérita Universidad Autónoma de Puebla

Facultad de Ciencias de la Computación

Algoritmos Genéticos para el Agrupamiento de datos de la Calidad del Aire.

Tesis de Licenciatura

Que para obtener el título de:
Licenciado en Ciencias de la Computación

Presenta:
Angel Omar Mendoza Rojas

Asesor de la tesis:
Dr. Abraham Sánchez López

Puebla, Pue. **Verano 2012**

RESUMEN

Los algoritmos genéticos están basados en la teoría evolutiva de Carlos Darwin, en este trabajo se presentaran algunos conceptos que permitirán entender el comportamiento de algunos factores que son indispensables para el mejor funcionamiento de los algoritmos genéticos como: el operador de selección, operador de mutación, operador de cruza. El método que se implementó es el agrupamiento por racimos de semillas (CSPM, sus siglas en ingles), que ha demostrado una mayor eficiencia y eficacia en el tiempo de respuesta de un resultado (no necesariamente el óptimo), usando datos del cambio climático que permitirán mostrar las características de nuestra metodología y el comportamiento de estos datos con el algoritmo genético.

AGRADECIMIENTOS

A mis abuelos, padres, y hermanas que fueron pilares y motor del esfuerzo de cada día, durante este largo trayecto. Reconozco su apoyo, esfuerzo y sacrificio incondicional que tuvieron hacia mi persona.

Una mención muy especial y respetuosa a todas aquellas personas que contribuyeron en mi formación académica, en mi persona y que lograron gratos recuerdos en mi estancia universitaria. Por su paciencia, compañerismo y su amistad.

Rubén Diario Rojas Bravo, Dolores León Flores, Domingo Miguel Mendoza Mendoza, Ma. Julia Yolanda Rojas León, Norma Yolanda Mendoza Rojas, Marisol Mendoza Rojas, Francisco Javier Rojas León, Dr. Abraham Sánchez López, Juan Carlos Conde Ramírez, Juan Carlos Pérez Medina, Miguel Díaz Ruiz, Kevin Omar García Manzo, Dalia Rodríguez Salas, Marco López Ortiz.

Muchas gracias.

ÍNDICE

ÍNDICE DE TABLAS

ÍNDICE DE EJEMPLOS

ÍNDICE DE FIGURAS

I. INTRODUCCIÓN

El clustering (“Segmentación”), también llamada agrupamiento, permite la identificación de tipologías o grupos donde los elementos guardan gran similitud entre sí y muchas diferencias con los de otros grupos. Así, se puede segmentar el colectivo de clientes, el conjunto de valores e índices financieros, el espectro de observaciones astronómicas, el conjunto de zonas forestales, el conjunto de empleados y de sucursales u oficinas, etc. La segmentación está teniendo mucho interés desde hace ya tiempo dadas las importantes ventajas que aporta al permitir el tratamiento de grandes colectivos de forma pseudo-particularizada, en el más idóneo punto de equilibrio entre el tratamiento individualizado y aquel totalmente masificado.

El clustering es el proceso mediante el cual se realiza un proceso de clasificación sobre una población de elementos determinada, de manera tal que la clasificación se realice un ordenamiento de comportamientos similares de forma tal que la selección sea de un grado alto para elementos de determinado grupo y baja para elementos de grupos diferentes. La importancia que hoy en día tienen los algoritmos genéticos se debe en gran medida a la eficiencia que han demostrado en distintos campos de estudio y que nos permiten presentar posibles soluciones para ciertos problemas, por otro lado el clustering no ha tenido un avance tan significativo, ha sido con paso más lento.

Por eso la importancia de seguir trabajando sobre el desarrollo de clustering. Este proceso llega a requerir un gran número de recursos computacionalmente hablando, aunque existen metodologías para garantizar un número exitoso de clustering dependiendo del tamaño del problema, no hemos considerado definir un número base de clustering para este trabajo de investigación, es decir, nuestro clustering dependerá su eficiencia en el número de clusters que considere necesario en base a su aprendizaje para la resolución de problemas.

Algunas metodologías que se podrían considerar para el desarrollo de clusters son: algoritmo de Calinski y K-means, este último es el más utilizado. En los años 60's Lawrence J. Fogel propuso una técnica denominada "Programación Evolutiva", en la cual la inteligencia se ve como un comportamiento adaptativo y enfatiza los nexos de comportamiento entre padres e hijos, en vez de buscar emular operadores genéticos específicos (como los Algoritmos Genéticos).

El algoritmo básico de la Programación Evolutiva es el siguiente:

- Generar aleatoriamente una población inicial.
- Se aplica mutación.
- Se calcula la aptitud de cada hijo y se usa un proceso de selección mediante torneo (normalmente estocástico) para determinar cuáles serán las soluciones que se retendrán.

Algunas aplicaciones de la Programación Evolutiva son:

- Predicción.
- Generalización.
- Juegos.
- Control automático.
- Problema del agente viajero.
- Planeación de rutas.
- Diseño y entrenamiento de redes neuronales.
- Reconocimiento de patrones.

Una de las ramas más desarrolladas en el área de la Inteligencia artificial son los algoritmos genéticos que son métodos adaptativos empleados para la resolución de ciertos problemas, utilizando información histórica para encontrar nuevos puntos de búsqueda que nos lleven a una solución optima de determinado problema. Los Algoritmos genéticos son una técnica de búsqueda basada en la teoría de la evolución de Darwin, que ha cobrado tremenda popularidad alrededor del mundo durante los últimos años [5].

Por imitación de este proceso, los Algoritmos genéticos son capaces de ir creando soluciones para problemas del mundo real [4]. La evolución de dichas soluciones hacia los valores óptimos del problema depende en buena medida de una adecuada codificación de las mismas. No podemos dejar a un lado que el soft computing es sin duda alguna, una técnica innovadora en el creación de sistemas inteligentes, donde su principal desarrollo consiste en sistemas robustos para obtener soluciones aceptables en la resolución de un problema en especifico. Los principales componentes del soft computing son la lógica difusa, las redes neuronales, el computo evolutivo, el aprendizaje de maquinas y el razonamiento probabilístico.

A pesar de que se tienen estudios desde los años 20 sobre los algoritmos genéticos y algunas importaciones indispensables hasta hoy, como las de Fogel en los años 60, es a partir de los años 70 como los algoritmos genéticos toman una importancia, y día a día han mejorado en su desempeño y tienen relevancia en su estudio, basados en la adaptación de las especias en un entorno dado, es decir en su evolución.

El cambio climático llega hacer abrupto, y en general está asociado directa o indirectamente a la actividad humana. La razón principal del incremento de la temperatura se debe al proceso de industrialización y en algunos casos es producida por grandes cantidades de combustión de origen fósil como petróleos y carbón, a la tala desmedida de bosques y métodos inadecuados de explotación agrícola. Estas actividades han aumentado las emisiones de gases de efecto invernadero (GEI) en la atmosfera, sobretodo de metano, bióxido de carbono y oxido nitroso. El exceso de estos gases produce en forma desmedida elevadas temperaturas y modifica las condiciones climáticas. Tan solo en los últimos 15 años la temperatura media de la superficie terrestre ha incrementado en 0.6°C.

A mediados del siglo XX, el crecimiento demográfico y económico en la ciudad de México derivo en una expansión física, que provoco el desbordamiento de sus límites sobre los municipios periféricos del Estado de México, dando lugar al proceso de metropolización, teniendo como características la falta de planeación y regulación del desarrollo urbano. La cuenca del Valle de México situada a 2,240 metros de altura sobre el nivel del mar se encuentra rodeada por una cadena montañosa integrada por la Sierra de Monte Bajo, Sierra

de las Cruces, Sierra del Chichinautzin, Sierra Nevada, Sierra del Río Frío. Entre los principales factores fisiográficos y climáticos que afectan la calidad del aire de la cuenca destacan los siguientes:

El entorno montañoso que la rodea, ya que constituye una barrera natural que dificulta la libre circulación del viento y la dispersión de los contaminantes. Por su altitud, frecuentemente ocurren inversiones térmicas en el Valle en un importante porcentaje de los días del año. Éste es un fenómeno natural que causa un estancamiento temporal de las masas de aire en la atmósfera.

La intensa radiación solar que se registra en el Valle de México durante todo el año favorece la formación del ozono, lo cual provoca que la luz ultravioleta del sol desencadene entre los óxidos de nitrógeno y los hidrocarburos emitidos a la atmósfera, los cuales son precursores del ozono y junto con los óxidos de azufre precursores de partículas finas.

En los años 80, en medio de la crisis económica de larga duración y del ajuste estructural neoliberal, la configuración territorial de la zona metropolitana del Valle de México (ZMVM) estuvo sometida al proceso acelerado de cambio, caracterizado por un patrón de crecimiento y estructuración urbana regido por la iniciativa privada, el libre mercado, la desregulación y el debilitamiento de la política estatal. Los municipios conurbados, con alto crecimiento poblacional, se constituyeron como áreas de localización industrial, de habitación de capas medias y sobre todo de sectores populares, que han servido en gran medida como zonas-dormitorio. El crecimiento, continúa hasta ahora, integrando un número cada vez mayor de municipios y degradando sus reservas naturales y áreas rurales.

La calidad del aire en diversas ciudades de México se ha deteriorado significativamente, la mayor parte de los procesos de urbanización y de crecimiento poblacional, así como, de las actividades económicas se han dado en ausencia de una reglamentación y de programas específicos para enfrentar los diversos problemas ambientales que padecen las ciudades mexicanas, particularmente la contaminación atmosférica.

La calidad del aire es producto de una combinación de factores naturales y sociales. Los factores climatológicos y geográficos constituyen elementos que agravan u obstaculizan la

solución de la contaminación del aire. No obstante, la causa principal del deterioro de la calidad del aire son las actividades humanas, especialmente las de carácter económico.

En algunas zonas determinadas, la calidad del aire puede verse afectada por elementos climáticos y geográficos, está relacionada directamente con el volumen y características de los contaminantes emitidos de forma local y regionalmente a la atmósfera. Por ello, un componente indispensable para el diseño y la aplicación de cualquier programa para controlar el problema de la contaminación del aire es la información sobre las principales fuentes de contaminantes atmosféricos y los volúmenes emitidos.

Los contaminantes entran en contacto con el aire mediante fuentes naturales o sintéticas. El aire siempre porta contaminantes naturales como polen, esporas, moho, levaduras, hongos y bacterias; los incendios forestales, los vendavales, las erupciones volcánicas y las sequías producen humo, aerosoles y otros contaminantes que entran al aire. La contaminación que surge de la naturaleza encuentra poca comparación con los efectos de los contaminantes asociados con las actividades humanas.

Como dato interesante en 1999 se emitieron 40.5 millones de toneladas de contaminantes atmosféricos (58% por fuentes naturales y 42% por fuentes antropogénicas, es decir, que son producidos por el hombre). En cualquiera de sus formas, se deriva en alteraciones del ambiente, en el caso de la contaminación atmosférica, es “la presencia en el aire de toda materia o energía en cualquiera de sus estados físicos y formas, que al incorporarse o actuar en la atmósfera altera o modifica su composición y condición natural” (Sedue, 1989: 1).

El Índice metropolitano de la calidad del aire (IMECA) es una unidad mediante la cual se relaciona la concentración de contaminantes con los límites establecidos por las normas y con sus efectos sobre la salud. El IMECA brinda las mediciones diarias por hora y por zona de la concentración del ozono (O3), bióxido de nitrógeno (NO2), bióxido de azufre (SO2), monóxido de carbono (CO) y partículas menores a 10 μm (PM 10), según las estaciones de monitoreo. En aquellas ocasiones en que los valores sean altos y se mantengan por largos periodos de tiempo, con el fin de mantener la salud de la población se aplica el “Programa de Contingencias”.

Por lo anterior, se consideró que es importante el estudio sobre los procesos de clustering que nos permitan tener una herramienta para el desarrollo de nuevas aplicaciones. Usamos algunas características de la programación evolutiva, así como del soft computing y de redes neuronales para procesar algunos datos sobre el cambio climático y una forma de agrupar ciertas coincidencias de los datos en clusters.

Los capítulos en los que está basado el trabajo de investigación son acerca del cambio climático, del proceso de clustering (o agrupamiento) y algoritmos genéticos, la implementación de algoritmos genéticos, nuestros resultados y aplicaciones, y finalmente las conclusiones y trabajo a futuro.

II. CAMBIO CLIMÁTICO

En México cada día va tomando mayor importancia tener conocimiento de datos que nos permitan tener una certeza sobre la alteraciones que influyan en el cambio climático. México ha ratificado el tratado de Kioto y en los últimos años, los organismos gubernamentales han integrado planes en combate al cambio climático. Algunos de los impactos económicos en que se deriva las afectaciones son palpables en la escasa de bienes y servicios, repercusiones en el suministro de energía, transporte y distribución de la infraestructura. Y se prevé que en los próximos 20 años el apoyo ambiental que se otorgue a la industria y la agricultura sea más escaso, tornando a la producción y la industria manufacturera que encarezcan sus costos. La variabilidad del clima está presente en México, de acuerdo con la información del Servicio Meteorológico Nacional (SMN), se extenderán los periodos de sequía provocando temperaturas extremas en parte del país; no sólo en zonas áridas y semiáridas del norte y noreste, se incluyen también regiones del sur que históricamente han sido húmedas.

La interacción entre los impactos del cambio climático y las consecuencias en materia de seguridad es compleja, es posible evaluar algunos efectos de seguridad en México. Un ejemplo es, que se prevé que en la medida que se altere la disponibilidad del agua y la producción de alimentos, los estilos de vida sufrirán repercusiones; provocarán el incremento de movilidad poblacional que intente emigrar en busca de mejores condiciones (ya sea dentro del país o en el extranjero). Asimismo, se considera que zonas de alto riesgo de sequia en México, aumentará el total de días secos consecutivos y, por tanto, las ondas de calor serán más extensas. Por otro lado existen numerosas ciudades que resultarán afectadas, por la elevación de los niveles de mar que predice el Grupo Intergubernamental de Expertos sobre el Cambio Climático (IPCC, sus siglas en inglés). En México las regiones más afectadas bajo estas condiciones son Cancún, Cozumel Veracruz e Ixtapa. El cambio climático es también un fenómeno que crea las condiciones necesarias para la construcción de nuevos mercados, nuevos modelos de producción y consumo, nuevos empleos y el desarrollo de nueva tecnología.

III. CLUSTERING Y ALGORITMOS GENETICOS

3.1 CLUSTERING.

El clustering surge con la intención de combinar diferentes procesos entre varias computadoras, para después recoger los resultados que debían producir, donde se pretende que conforme sea la grandeza del problema, el clustering se expanda producto de las necesidades que vayan aumentando.

El análisis de conglomerados o agrupamiento (*clustering*) es una técnica multivariante que busca agrupar elementos (o variables) tratando de lograr la máxima homogeneidad en cada grupo y la mayor diferencias entre el resto de grupos. Así pues, podríamos dar una pequeña introducción a la definición de cluster como un conjunto de computadoras interconectadas con dispositivos de alta velocidad que actúan en conjunto usando el poder de cómputo de varias CPUs en combinación para resolver ciertos problemas dados. [19]

El clustering es un proceso que permite agrupar o clasificar a diferentes elementos según las propiedades que compartan, es decir que el objetivo del clustering es el de ordenar las observaciones en grupos, de forma tal que el grado de asociación natural sea alto entro los miembros de un cierto grupo. El clustering nos permite contar con algunas características interesantes que nos darán la oportunidad de incrementar la escabilidad, disponibilidad y fiabilidad:

Escabilidad: es la capacidad de un equipo de hacer frente a volúmenes de trabajo cada vez mayores, sin dejar por ello de prestar un nivel de rendimiento aceptable.

Disponibilidad: es tener la capacidad que en determinado momento se haga uso de los recursos computacionales necesarios para su ejecución.

Fiabilidad: es la probabilidad de funcionamiento correcto (sin que esto signifique el resultado obtenido sea necesariamente el que resuelva el problema).

Construir un cluster tiene importantes ventajas y una gran variedad de aplicaciones:

- Incremento de velocidad de procesamiento ofrecido por clusters de alto rendimiento.
- Incremento en el número de transacciones o velocidad de respuesta ofrecida por los clusters de balanceo de carga.
- Incremento de la confiabilidad y la robustez ofrecido por los clusters de alta disponibilidad [18].

Las herramientas de segmentación se basan en técnicas de carácter estadístico, de empleo de algoritmos matemáticos, de generación de reglas y de redes neuronales para el tratamiento de registros.

Para otro tipo de elementos a agrupar o segmentar, como texto y documentos, se usan técnicas de reconocimiento de conceptos. Esta técnica suele servir de punto de partida para después hacer un análisis de clasificación sobre los *clusters*. La principal característica de esta técnica es la utilización de una medida de similaridad que, en general, está basada en los atributos que describen a los objetos, y se define usualmente por proximidad en un espacio multidimensional. Para datos numéricos, suele ser preciso preparar los datos antes de realizar datamining sobre ellos, de manera que en primer lugar se someten a un proceso de estandarización.

El cálculo de un k óptimo puede resultar innecesario, si es posible seleccionar un subconjunto representativo. Existe un muestreo llamado, *"muestreo estadístico"*, que es empleado para obtener una interpretación de una población sin necesidad de recabar información sobre el total de la población, sino sobre una muestra representativa. La forma de selección de muestra y su tamaño se basan en dos actores principales; podemos conocer el tamaño que tendrá la muestra, fijando el error de la muestra (1), el error de muestra se define como la raíz cuadrada de la varianza:

$$e(x) = \sqrt{\sigma(x)} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - x')^2} \quad (1)$$

Donde:

N = numero toral de individuos que componen la población.

x_i = valor de la variable x para el elemento i.

$x' = \frac{1}{N}\sum_{i=1}^{N}(x_i - x')^2$, la media

Una vez fijado el error demuestra, que podemos conocer el tamaño que tendrá la muestra aplicando la formula:

$$n = \frac{N * S^2}{N * e^2 + S^2} \quad (2)$$

Donde:

n = tamaño de la muestra.

N = número total de individuos que componen la población.

e = error de muestreo (calculado con la fórmula anterior).

S = quasivarianza = $\sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - x')^2}$

A continuación se describe los métodos de agrupamiento utilizados en este trabajo de tesis.

3.1.1 MÉTODO DE AGRUPAMIENTO JERÁRQUICO.

El método de agrupamiento jerárquico parte en un principio de la existencia de dos clusters como objetos, estos grupos se combinan de acuerdo al grado de mejora de sus valores objetivos y todos los subgrupos son fusionados en un solo grupo [33]. A continuación se presenta el algoritmo del método de agrupación jerárquica, para agrupar N objetos formulados como un problema de maximización.

Paso 0. Se inicia con N grupos, cada grupo tiene un único objeto, es decir:

$$S_i^{(1)} = \{O_i\}, \text{ i} = 1, \ldots, \text{N}$$

Una matriz simétrica de N x N es la función objetivo incremental,

$$\text{MF} = \{\Delta F_{ij}, \text{ i, j} = 1, \ldots, \text{N, i} \neq \text{j}\},$$

Donde ΔF_{ij} representa el incremento del valor objetivo en caso de que el i-esimo cluster y el j-esimo cluster sean combinados en un solo clusters. Tal que k = 1.

Paso 1. Si $\Delta F_{uv} = \max\{\Delta F_{ij}, \text{ i, j} = \text{l}, \ldots, \text{N - k, i} \neq \text{j}\}$ y v > u, entonces

$$S_u^{(k+1)} = S_u^{(k)} \, U \, S_v^{(k)},$$

$$S_1^{(k+1)} = S_1^{(k)}, \ldots, S_{u-1}^{(k+1)} = S_{u-1}^{(k)},$$

$$S_{u+1}^{(k+1)} = S_{u+1}^{(k)}, \ldots, S_v^{(k+1)} = S_{v+1}^{(k)}, \ldots, S_{N-v-1}^{(k+1)} = S_{N-v}^{(k)}$$

$$\text{Y } S_{N-k}^{(k+1)} = \Phi.$$

Calcula el valor objetivo de $F(X)^{(k)}$ de la partición. Tal que k = k + 1.

Paso 2. Repetir el paso 1 hasta que k = N - 1. F(X)* = max {$F(X)^{(k)}$, k = 1, . . . , N - 1}.

3.2.2 MÉTODO DE AGRUPAMIENTO SIMULTÁNEO

Este método se debe de garantizar la existencia de por lo menos dos objetos para que de esta forma se pueda formar un nuevo cluster, el número máximo de clusters debe ser de la forma K = [N/2]. Entonces habrá un total de N x K variables de decisión del problema [CA], así como lo muestra la tabla 3.1.

Tabla3.1. Relación ente N objetos y K clusters.

Objeto	Clusters					
	1	2	…	K	…	K
1	X_{11}	X_{12}		X_{1k}		X_{1K}
2	X_{21}	X_{22}		X_{2k}		X_{2K}
…						
i	X_{i1}	X_{i2}		X_{ik}		X_{iK}
…						
N	X_{N1}	X_{N2}		X_{Nk}		X_{NK}

Si existen más de dos variables con un valor de 1 en la misma columna significa que entre sí forman un mismo grupo. Sin embargo si X_{ik} es codificado como un gen, la longitud del cromosoma será demasiado largo y dará como resultado una insuficiencia de memoria de la computadora. Además, se dificultará el manejo de las restricciones, si una y sólo una variable es igual a 1 y también es igual a 0 en la misma fila. El método de agrupamiento

simultáneo (SICM) utiliza una técnica de codificación y decodificación para sustituir cada fila de la matriz de variables de decisión con un gen más corto de la cadena.

Por ejemplo, si se tienen 17 objetos, entonces hay al menos 8 particiones (8 = [17 / 2]). De modo tal que cada objeto pueda ser asignado a cualquier conjunto, estos grupos requieren de tres genes para representarlos, como se muestra en la Tabla 3.2.

Tabla 3.2.Reglas de coincidencia de cadenas de genes, enteros y clusters.

Cadenas de genes	Enteros	Clusters
000	0	1
001	1	2
010	2	3
011	3	4
100	4	5
101	5	6
110	6	7
111	7	8

Se remplaza cada fila con tres genes, esto no sólo reduce la longitud de los cromosomas (cuatro genes puede representar el problema de los 33 objetos, cinco genes puede representar el problema de los 65 objetos), sino también evita que en el problema, un objeto puede ser asignado o no a varios grupos. La tabla 3.3 ilustra el resultado de la agrupación factible de los 17 objetos.

El cromosoma de la Tabla 3.3 se compone de 51 genes (000001011000101101010000001000000011011000000101010). Entonces cada tres genes del cromosoma son decodificadas en un número entero de 0 - 7 de forma secuencial, lo que representa la agrupación de cada objeto y se le asigna de acuerdo a las normas de congruencia que se indican en la Tabla 3.2. Después de ser decodificados, los cromosomas

representan que cinco clusters se han formados. Los grupos se componen de 6, 2, 2, 3, 3 objetos, respectivamente.

Tabla 3.3. Relación entre objetos y clusters (17 objetos por ejemplo).

Objetos	Clusters								Decodificación
	1	2	3	4	5	6	7	8	
1	1	0	0	0	0	0	0	0	000
2	0	1	0	0	0	0	0	0	001
3	0	0	0	1	0	0	0	0	011
4	1	0	0	0	0	0	0	0	000
5	0	0	0	0	0	1	0	0	101
6	0	0	0	0	0	1	0	0	101
7	0	0	1	0	0	0	0	0	010
8	1	0	0	0	0	0	0	0	000
9	0	1	0	0	0	0	0	0	001
10	1	0	0	0	0	0	0	0	000
11	1	0	0	0	0	0	0	0	000
12	0	0	0	1	0	0	0	0	011
13	0	0	0	1	0	0	0	0	011
14	1	0	0	0	0	0	0	0	000
15	1	0	0	0	0	0	0	0	000
16	0	0	0	0	0	1	0	0	101
17	0	0	10	0	0	0	0	0	010
Subtotal	6	2	2	3	0	3	0	0	

3.2.3 MÉTODO DE AGRUPACIÓN POR ETAPAS.

El método de agrupación por etapas (STCM, sus siglas en inglés) sucesivamente resuelve la agrupación binaria óptima de un grupo hasta que el valor objetivo no pueda ser mejorado.

Un grupo inicial únicamente contiene todos los objetos que se dividen en dos subgrupos de tal manera que la función objetivo sea optimizada en esta etapa.

A través de cada proceso binario se agrupa, cada grupo se divide en dos subgrupos. Un grupo será nombrado como "sondeado", cuando no pueden estar agrupados en más grupos binarios para mejorar el valor objetivo. Este concepto es similar al de "branch and bound". El método de agrupación por etapas alcanzará la agrupación óptima, cuando todos los clusters presentan el estado de "sondeados". El algoritmo del modelo se muestra en la fig. 1. Los siguientes pasos son el algoritmo para el modelo de agrupación por etapas (STCM) en profundidad, que explora cada rama individualmente a la vez.

Paso 0. Sea $S^{(0)}$ que representan el grupo que contiene todos los objetos. El problema de la división óptima $S'^{(0)}$ en dos subgrupos, a saber, $S'^{(0)}$ y $S''^{(0)}$, se puede formular de la siguiente manera 0 - 1 programación matemática:

$$MP^{(0)}$$
$$Max\ F(X)^{(0)}$$
$$sujeto\ a\ X_1 = \{0,1\}\ i = 1, \dots, |S^{(0)}|$$

Donde $X_i = 1$ denota que i-ésimo objeto de $S'^{(0)}$ se agrupan en el grupo $S'^{(0)}$, $X_i = 0$ denota que el i-ésimo objeto de $S^{(0)}$ se agrupa en el grupo $S''^{(0)}$. X_i se codifica como gen de los cromosomas (la longitud de los cromosomas es $|S^{(0)}|$), y luego los algoritmos genéticos son empleados para resolver MP $^{(0)}$ al maximizar $F(X)^{(0)}$ para lograr la agrupación binaria óptima: $S'^{(0)} = \{ O_i \mid X_i^* = 1, O_i \in S^{(0)}\}$ y $S''^{(0)} = \{ O_i \mid X_i^* = 0, O_i \in S^{(0)}\}$

Paso 1. Sea $S^{(1)} = S'^{(0)}$ y cambiar la numeración de los objetos de $S^{(1)}$. Formular la agrupación binaria óptima de $S^{(1)}$como $MP^{(1)}$, que también se resuelve por los algoritmos genéticos. $f(X)^{(1)*}$ es el valor objetivo del grupo binario óptimo $S^{(1)}$ bajo el supuesto de que el otro ($S''^{(0)}$) no se modifica. El resultado agrupado es: $S'^{(1)} = \{ O_i \mid X_i^* = 1, O_i \in S^{(1)}\}$ y $S''^{(1)} = \{ O_i \mid X_i^* = 0, O_i \in S^{(1)}\}$. Tres grupos son formados, que son: $S''^{(0)}$, $S'^{(1)}$ y $S''^{(1)}$. $F(X)^{(1)*}$ es el valor objetivo de estos tres grupos.

Paso 2. Sea $S^{(i)} = S'^{(i-1)}$ y resolver $MP^{(i)}$ por un algoritmo genético.

Paso 3. Repetir el paso 2 hasta $S''^{(k)} = \varphi$, entonces esta rama es sondeada. Esto es un total de k+1 grupos, es decir, $S''^{(0)}$, $S''^{(1)}$, . . ., $S^{(k-1)}$ y $S'^{(k)}$. $S'^{(k)}$ ya que no puede ser dividido en los siguientes pasos. $F(X)^{(k)}*$ es el valor objetivo óptimo de estos grupos k + 1.

Paso 4. Elegir una de las ramas para que queden agrupadas binariamente repitiendo los pasos 2 y 3 hasta que sea sondeada.

Paso 5. Si todas las ramas son sondeadas, y luego se detiene. Los grupos formados son la agrupación óptima, resultado del STCM. De lo contrario, ir al paso 4.

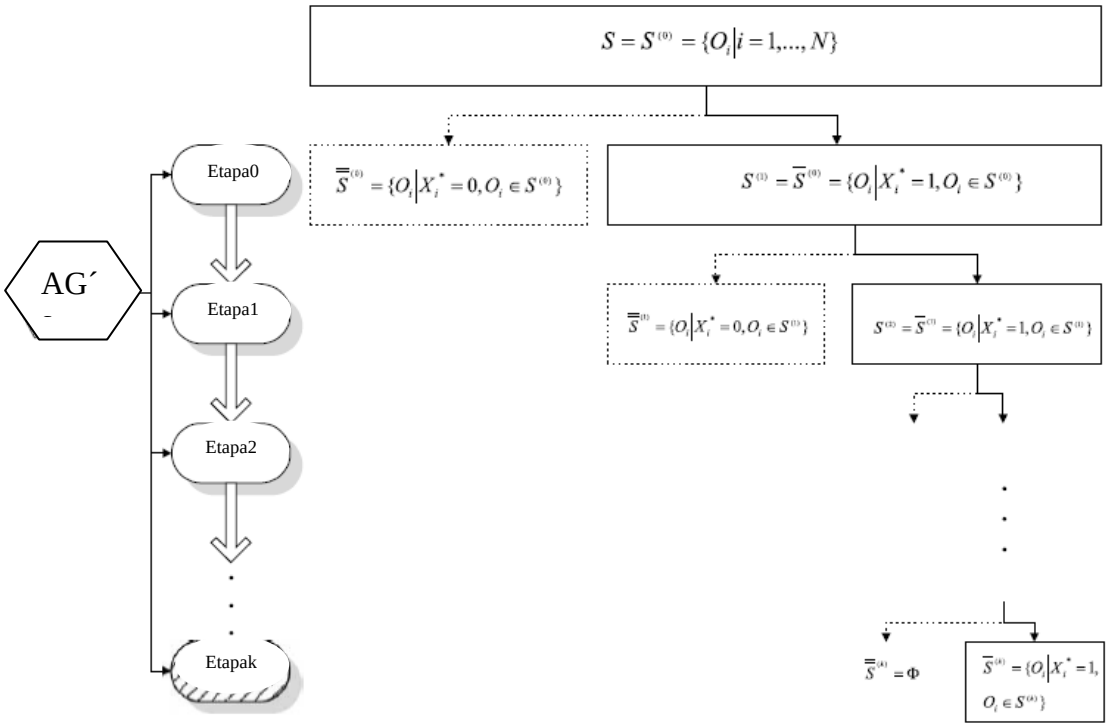

Figura 3.1. Framework de STCM.

En comparación al método de agrupamiento simultaneo (STCM), la codificación y decodificación del método de agrupamiento por etapas (STCM) son mucho más simples debido a que la longitud del cromosoma puede ser reducido en gran medida y además tiene

la posibilidad de reducirse aún más en la evolución de cada etapa de optimización. Consideremos también la posibilidad de N objetos, por ejemplo. Sea $|S^{(0)}| = N$, denotando N objetos en el conjunto $S^{(0)}$, la longitud del cromosoma en la etapa 0 es N. Si $| S''^{(0)} | = L_0$, la longitud del cromosoma en la etapa 1 es $N - L_0$. Si $S''^{(1)} = L_1$, la longitud del cromosoma en la fase 2 puede ser más cortas, como $N - L_0 – L_1$, y así sucesivamente.

3.2.4 MÉTODO DE AGRUPAMIENTO POR PUNTOS DE SEMILLAS.

El método de agrupamiento por puntos de semillas (CSPM, sus siglas en inglés) utiliza en primer lugar a los algoritmos genéticos para realizar la selección de las semillas del grupo más adecuado de todos los objetos, entonces se toma como muestra. El resto de los objetos de cada grupo será tomado en cuenta de acuerdo a la similitud con la semilla del grupo o de su grado de mejoría con respecto a la función objetivo. El número de semillas por grupo representa el número de cada uno de los grupos y las características de estas semillas determinan el racimo de la agrupación resultado. El marco del método de agrupamiento por puntos de semilla (CSPM) está representado en la figura 3.2.

Los siguientes, son los pasos del algoritmo de asignación de la figura 3.2.

Paso 0. Sea k=1 y sea S un conjunto de todos los objetos, que es $S=\{O_1, . . ., O_N\}$. CP_m es un conjunto de grupos de semillas, donde, $CP_m = \{c_1, . . ., c_m\}$. NP es un conjunto de semillas no agrupadas, es decir, $NP = S - CP_m \cdot S_j^{(0)} = \{c_j\}$, j = 1,. . ., m.

Paso 1. Sea O_k, donde k es el k-ésimo objeto de NP. Si $F(S_1^{(k)}, . . ., S_{j-1}^{(k)}, S_j^{(k)} U \{O_k\}, S_{j+1}^{(k)}, . . ., S_m^{(k)}) = Max_i \{ F(S_1^{(k)}, . . ., S_{i-1}^{(k)}, S_i^{(k))} U \{O_k\}, S_{i+1}^{(k)}, . . ., S_m^{(k)} \}$, entonces O_k es asignado al j-ésimo grupo.

Paso 2. Sea $S_j^{(k)} = S_j^{(k-1)} U \{O_k\}$ y k = k+1. Si $k < N – m + 1$, regresar al paso 1, en caso contrario terminar.

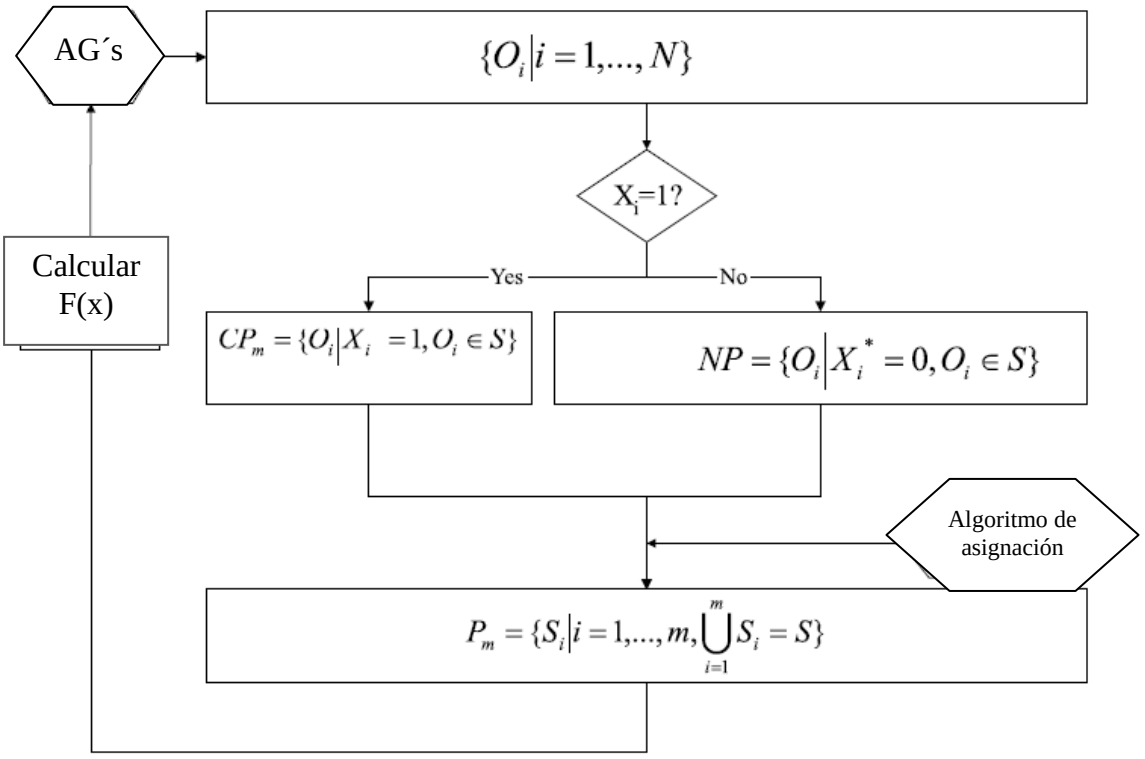

Figura 3.2. Framework de CSPM.

Una vez que el resultado de la agrupación, de P_m, se obtiene, el valor objetivo F(X) que representa la adecuación de este conjunto de grupo de semillas, también se ha determinado. Sin embargo, CSPM emplea algoritmos genéticos para buscar las semillas en el racimo optimo por variables como los genes que codifican X_i para representar los objetos relacionados, donde $X_i = 1$ denota que el i-ésimo objeto se elige como una semilla del racimo, y $X_i = 0$ en caso contrario.

3.2 ALGORITMOS GENÉTICOS.

Los Algoritmos Genéticos (AGs) son una técnica que tiene su inspiración, en la Biología, al igual que las Redes Neuronales. Estos algoritmos representan el modelado matemático de como los cromosomas en un marco evolucionista alcanzan la estructura y composición más

óptima en aras de la supervivencia. Entendiendo la evolución como un proceso de búsqueda y optimización de la adaptación de las especies que se plasma en mutaciones y cambios de modelos de reproducción (mutación y cruza) para ser utilizadas en todo tipo de problemas de búsqueda y optimización. Se da la mutación cuando alguno o algunos de los genes cambian bien de forma aleatoria o de forma controlada vía funciones y se obtiene la cruza cuando se construye una nueva solución a partir de dos contribuciones procedentes de otras soluciones "padre". En cualquier caso, tales transformaciones se realizan sobre aquellos especimenes o soluciones más aptas o mejor adaptadas.

Dado que los mecanismos biológicos de evolución han dado lugar a soluciones (en el caso de los seres vivos), realmente idóneas, cabe esperar que la aplicación de tales mecanismos a la búsqueda y optimización de otro tipo de problemas tenga el mismo resultado. De esta forma los AGs transforman los problemas de búsqueda y optimización de soluciones en un proceso de evolución de unas soluciones de partida.

Las soluciones se convierten en cromosomas, transformación que se realiza pasando los datos a formato binario (aunque hay otros tipos de formato), y a los mejores se les van aplicando las reglas de evolución (funciones probabilísticas de transición) hasta encontrar la solución óptima. En muchos casos, estos mecanismos brindan posibilidades de convergencia más rápidos que otras técnicas.

Todos los organismos que conocemos están compuestos por una o más células, cada una de las cuales contiene a su vez uno o más cromosomas, esto es, cadenas de ADN. Un cromosoma se puede dividir, conceptualmente, en genes, bloques funcionales de ADN que codifican una determinada proteína. Solemos pensar que los genes solo son los - responsables de determinar los rasgos de un individuo, como el color de los ojos, o del cabello. Los AG son métodos adaptativos que pueden usarse para resolver problemas de búsqueda y optimización. Están basados en el proceso genético de los organismos vivos. A lo largo de las generaciones, las poblaciones evolucionan de acuerdo con los principios de la selección natural y la supervivencia del más fuerte, de acuerdo a los postulados de Darwin (1859). Por imitación de este proceso, los AG son capaces de ir creando sus propias soluciones para problemas del mundo real. La evolución de dichas soluciones dirigida a los

valores más óptimos del problema, depende en gran medida de una buena codificación de las mismas.

John Holland y sus colegas de la Universidad de Michigan generaron "algoritmos genéticos", donde su función estaba basada en dos objetivos, el primero consistía en abstraer y explicar rigurosamente el proceso adaptativo de los sistemas naturales, mientras que el segundo pretendía diseñar sistemas artificiales que retuvieran los mecanismos más importantes de los sistemas naturales. En este sentido, podemos decir que los algoritmos genéticos son *Algoritmos de búsqueda basados en los mecanismos de selección natural y la genética. En otros mecanismos combinan la supervivencia de los más compatibles entre las estructuras de cadenas, con una estructura de información aleatoria, intercambiada para construir un algoritmo de búsqueda con algunas de las capacidades de innovación de la búsqueda humana.*

El uso de estos algoritmos no está tan extendido como otras técnicas, pero van siendo cada vez más utilizado directamente en la solución de problemas, así como en la mejora de ciertos procesos presentes en otras técnicas. Así, por ejemplo, se usan para mejorar los procesos de adiestramiento y selección de arquitectura de las redes neuronales, para la generación e inducción de árboles de decisión y para la síntesis de programas a partir de ejemplos ("Genetic Programming").

Los algoritmos de agrupamiento tienen dos componentes comunes: Una función objetivo, la cual es utilizada para determinar la calidad de un agrupamiento dado y una estrategia de búsqueda para explorar el espacio de todos los posibles agrupamientos.

Algunos algoritmos tradicionales de agrupamiento presentan problemas cuando el espacio de búsqueda es grande con respecto al número de dimensiones, o no es métrico, o no es unimodal, o cuando la función objetivo es ruidosa. Muchos investigadores de la comunidad de los Algoritmos Genéticos han encontrado que si el espacio de búsqueda es grande y se conoce que no es perfectamente plano o unimodal, o si la función de aptitud es ruidosa y si la tarea requiere una solución cercana al óptimo global, o un óptimo global a ser encontrado, es decir que es suficiente hallar una buena solución, entonces un AG tiende a ser competitivo frente a otros métodos.

El algoritmo genético básico consiste en los siguientes pasos:

- Generar aleatoriamente una población inicial del problema.
- Calcular la aptitud de cada individuo.
- Seleccionar probabilísticamente individuos con base a la aptitud.
- Aplicar operadores genéticos (cruza y mutación) para la generación de la siguiente población.
- Ciclar los pasos anteriores hasta que se tenga cierta condición se satisfaga como posible solución.

Un algoritmo genético también tiene una serie de parámetros que se tienen que fijar para cada ejecución, por ejemplo el tamaño de la población que debe de ser suficiente para garantizar que se está generando una diversidad de posibles soluciones, además tiene que crecer más o menos con el número de bits del cromosoma, aunque nadie ha aclarado cómo tiene que hacerlo. Por supuesto, depende también de la computadora en la que se esté ejecutando. Y una condición de paro, lo más habitual es que la condición de paro sea la convergencia del algoritmo genético o un número prefijado de generaciones. La función de aptitud no es más que la función objetivo de nuestro problema de optimización. El algoritmo genético únicamente se maximiza, pero la minimización puede realizarse fácilmente utilizando el recíproco de la función maximizar (debe cuidarse, por supuesto, que el recíproco de la función no genere una división por cero).

Una característica que debe tener esta función es que debe ser capaz de "castigar" a las malas soluciones, y de "premiar" a las buenas, de forma que sean estas últimas las que se propaguen con mayor rapidez. La codificación más común de las respuestas es a través de cadenas binarias, aunque se han utilizado también números reales y letras. El primero de estos esquemas ha gozado de mucha popularidad debido a que es el que propuso originalmente Holland, y además porque resulta muy sencillo de implementar [5]. Sabiendo la aptitud de cada cromosoma, se procede a la selección de los que se cruzarán en la siguiente generación (presumiblemente, se escogerá a los "mejores"). Como se puede observar, utilizar los algoritmos genéticos pudiera verse como algo complicado debido a los elementos que se deben procurar para su implementación, de esta manera se presentan sus

ventajas y desventajas de utilizar estos métodos para buscar una posible solución a la resolución de problemas.

Ventajas

- No necesitan conocimientos específicos sobre el problema que intentan resolver.
- Operan de forma simultánea con varias soluciones, en vez de trabajar de forma secuencial como las técnicas tradicionales.
- Cuando se usan para problemas de optimización, al maximizar una función objetivo las soluciones resultan menos afectadas por los máximos locales (falsas soluciones) que las técnicas tradicionales.
- Resulta sumamente fácil ejecutarlos en las modernas arquitecturas masivas en paralelo.
- Usan operadores probabilísticos, en vez de los típicos operadores determinísticos de otras técnicas.

Desventajas

- Pueden tardar mucho en converger, o no converger en absoluto, dependiendo en cierta medida de los parámetros que se utilicen –tamaño de la población, número de generaciones, etc.-.
- Pueden converger prematuramente debido a una serie de problemas de diversa índole.

Los algoritmos genéticos comúnmente son aplicados para la solución de problemas de optimización, en donde han mostrado ser muy eficientes y confiables. Sin embargo, no todos los problemas pudieran ser apropiados para la técnica, y se recomienda tomar en cuenta las siguientes características del mismo antes de intentar usarla:

1. Que su espacio de búsqueda es decir que sus posibles soluciones deben de estar delimitadas por un rango.

2. Además se debe definir una función de aptitud que nos permita conocer que tan cercanos nos encontramos de un resultado esperado.

3. Los resultados deben permitirnos implementarse de manera sencilla en una computadora.

Esto quiere decir que los algoritmos genéticos serán utilizados para resolver problemas que tengas espacios de búsquedas discretos, sin embargo, estos algoritmos también pueden ser utilizados en aquellos problemas donde su espacio de búsqueda contengan un espacio de soluciones continuas con la restricción de que exista preferentemente un rango de soluciones pequeño.

El poder de los Algoritmos Genéticos proviene del hecho de que se trata de una técnica robusta, y pueden tratar con éxito una gran variedad de problemas provenientes de diferentes aéreas, incluyendo aquellos en los que otros métodos encuentran dificultades. Si bien no se garantiza que el Algoritmo Genético encuentre la solución optima del problema, existe evidencia empírica de que se encuentran soluciones de un nivel aceptable, en un tiempo competitivo respecto a otros algoritmos.

Los algoritmos genéticos básicos son los algoritmos genéticos más simples y son en los que se basan el resto de variantes. Su funcionamiento se basa en intentar imitar la evolución natural para buscar la mejor solución para un problema. Los algoritmos genéticos operan sobre un grupo de posibles soluciones al que se le llama población y que inicialmente es creado aleatoriamente.

A cada una de estas posibles soluciones se le llama individuo o cromosoma y se representa como una cadena de números, habitualmente bits (0 y 1). Los individuos tienen una longitud concreta y a cada uno de los números que conforman un individuo se le llama gen. Lógicamente, la población que es generada al principio aleatoriamente será una población de malas soluciones, pero precisamente se trata de que la población vaya mejorando por medio de la evolución natural hasta conseguir buenas soluciones.

Los operadores genéticos serán los encargados en definir el comportamiento de los datos en base a los parámetros que se les sean asignados, en gran medida el resultado de la resolución del problema está en dependencia de los operadores genéticos por lo cual es la importancia de conocer el comportamiento y su funcionalidad de cada uno. La evolución natural se basa en que las especies se vayan adaptando a su entorno. Del mismo modo, en los algoritmos genéticos, las soluciones han de ir adaptándose al problema para que, al final, tengamos una población de buenas soluciones.

La función de *fitness* es la función que mide lo adaptado que está un individuo al problema. Esta función será la que diga lo cerca que está una posible solución de la mejor solución del problema: cuanto mayor sea la función de *fitness* de un individuo, más cerca estará de la mejor solución. La función de *fitness* es la única función específica de cada problema de todas las funciones que realiza un algoritmo genético. Es, en cierto modo, la que define el problema dentro del algoritmo genético.

3.2.1. OPERADOR DE SELECCIÓN.

Este operador escoge cromosomas entre la población para efectuar la reproducción. Cuanto más capaz sea el cromosoma, más veces será seleccionado para reproducirse. En otras palabras es el encargado de transmitir y conservar aquellas características de las soluciones que se consideran valiosas a lo largo de las generaciones. Sin embargo, es necesario también incluir un factor aleatorio que permita reproducirse a individuos que aunque no estén muy bien adaptados, puedan contener alguna "información útil" para posteriores generaciones, con el objetivo de mantener así también una cierta diversidad en cada población.

En cada generación de individuos debe hacerse una selección de los individuos más adaptados para pasar sus genes a la siguiente generación, al igual que ocurre con el proceso de la selección natural. Para acercarse lo más posible a ésta última, la selección en un

algoritmo genético se hará de manera aleatoria, pero teniendo cada individuo una probabilidad de ser seleccionado proporcional a su función de *fitness*, es decir, a la bondad de sus genes. De este modo conseguimos que en general se vayan eligiendo a los mejores individuos, pero, al mismo tiempo, los que no son tan buenos también tienen posibilidades de pasar y aportar algún gen que haga falta para formar mejores individuos en la siguiente generación.

3.2.1.1. SELECCIÓN POR RULETA.

Es el usado por Goldberg en su libro "Genetic Algorithms in Search, Optimization, and Machine Learning". Este método es muy simple, y consiste en crear una ruleta en la que cada cromosoma tiene asignada una fracción proporcional a su aptitud. Sin que nos refiramos a una función de aptitud en particular, supongamos que se tiene una población de 5 cromosomas cuyas aptitudes están dadas por los valores mostrados en la tabla 3.4.

Tabla 3.4: Valores de ejemplo para ilustrar la selección por ruleta.

Cromosoma No.	**Cadena**	**Aptitud**	**% Total**
1	11010110	254	24.5
2	10100111	47	4.5
3	110110	457	44.1
4	1110010	914	18.7
5	11110010	85	8.2
Total		1037	100

Con los porcentajes mostrados en la cuarta columna de la tabla 3.4. podemos elaborar la ruleta (Figura 3.3). Esta ruleta se gira 5 veces para determinar qué individuos se seleccionarán. Debido a que a los individuos más aptos se les asignó un área mayor de la ruleta, se espera que sean seleccionados más veces que los menos aptos [5].

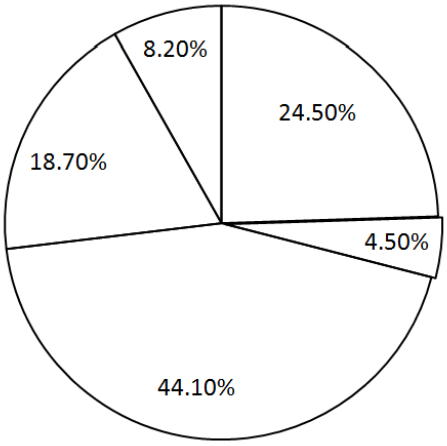

Figura 3.3. Representa los valores aptitud de la Tabla 3.4.

2.2.1.2. SELECCIÓN POR TORNEO.

La idea de este método es muy simple. La poblacion compite entre si (normalmente compiten en parejas, torneo binario) en un torneo del que y se selecciona un grupo de individuos que seran aquellos que tengan los valores mas altos de aptitud de la poblacion, al hacer un torneo binario se recomienda que la poblacion compita minimamente en dos ocasiones para garantizar que obtengamos a los mejores individuos (es decir, el mejor individuo sera seleccionado 2 veces).

Una vez realizada la selección de los individuos se procede a la reproduccion o cruza de individuos donde intercambiaran su material cromosomico que se vera mas adelatnte. Las dos formas más comunes de reproducción sexual son de un punto único de cruza y uso de dos puntos de cruza.

De hecho existe una técnica desarrollada hace algunos años en la que el individuo más apto a lo largo de las distintas generaciones no se cruza con nadie, y se mantiene intacto hasta que surge otro individuo mejor que él, que lo desplazará. Dicha técnica es llamada elitismo, y no debe sorprendernos el hecho de que haya sido desarrollada en Alemania. Además la selección y la cruza, existe otro operador, el cual nos permite introducir material cromosómico nuevo en la población, tal y como sucede con sus equivalentes biológicos.

Si supiéramos la respuesta a la que debemos llegar de antemano, entonces detener el algoritmo genético sería algo trivial. Sin embargo, eso casi nunca es posible, por lo que normalmente se usan 2 criterios principales de detención: ejecutar el algoritmo genético durante un número máximo de generaciones o detenerlo cuando la población se haya estabilizado, es decir, cuando todos o la mayoría de los individuos tengan la misma aptitud [5].

Los algoritmos genéticos enfatizan la importancia de una cruza sexual que es el operador principal, la mutación como un operador secundario y utiliza una selección probabilística.

3.2.2. OPERADOR DE CRUZA.

El operador de cruza nos permite realizar una exploración de toda la información almacenada hasta el momento en la población y combinarla para generar mejores individuos con respecto de la población anterior. Existen diversos métodos para realizar la operación de cruza. Una vez seleccionados los individuos cuyos genes pasarán a la siguiente generación, se procede a la formación de ésta.

Para ello primero se realizarán un número aleatorio de cruces (como máximo población/2). Los individuos que se vayan a cruzar también se deberán elegir aleatoriamente, teniendo en cuenta que un mismo individuo no puede cruzarse más de una vez. El cruce consiste en un intercambio de genes entre dos individuos a partir de un punto que, como casi todo, se elige aleatoriamente. A continuación se muestra un ejemplo:

110|01 110|11 => 010|11 010|01

Ejemplo 3.1. : Cruce

Después de realizarse todos los cruces tendremos un nuevo grupo de individuos que poseerán nuevas características con respecto a los anteriores.

3.2.2.1. CRUZA EN UN PUNTO.

Es el método de cruza más sencillo que consiste en seleccionar dos individuos donde se cortan sus cromosomas por un punto que es seleccionado de manera aleatoria para generar dos segmentos que serán diferentes entre cada unos de ellos: la cabeza y la cola. De esta forma se hace el intercambio de colas entre los individuos para generar los nuevos descendientes, así ambos descendientes heredan información genética de los padres.

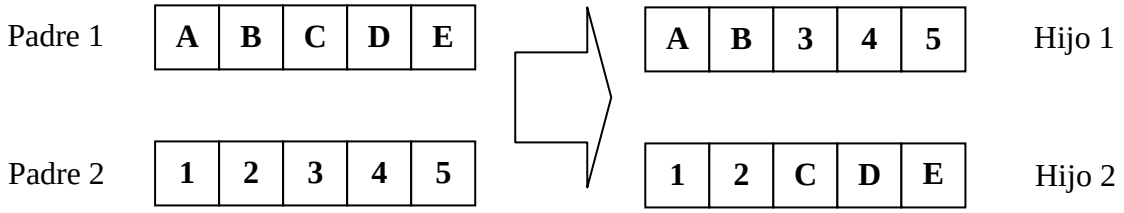

Figura 3.4. Cruza en un punto.

3.2.2.2. CRUZA EN N PUNTOS.

Es una generalización del método anterior. En este método en vez de cortar por un único punto los cromosomas de los padres como en el caso anterior se realizan dos cortes, donde deberá tenerse en cuenta que ninguno de estos puntos de corte coincida con el extremo de los cromosomas para garantizar que se originen tres segmentos, para generar la descendencia se escoge el segmento central de uno de los padres y los segmentos laterales del otro padre.

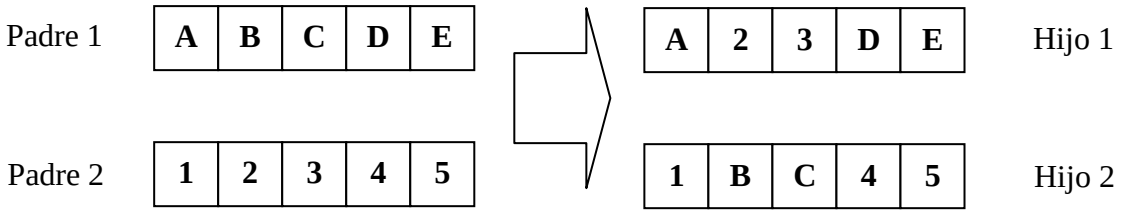

Figura 3.5. Cruza en dos puntos.

3.2.2.3. CRUZA UNIFORME.

Consiste en que cada gen tiene las mismas posibilidades de descendencia de pertenecer a uno u otro padre; esto se realiza de forma independiente al orden de codificación que impuso a cada uno dentro de su cromosoma, a pesar de que se puede implementar de maneras diversas, la técnica implica la generación de una máscara de cruce con valores binarios, si una de las posiciones de la máscara hay un 1, el gen situado en esa posición en uno de los descendientes se copia del primer padre, si por el contrario hay un 0 el gen se copia del segundo padre. Para producir el segundo descendiente se intercambian los papeles de los padres, o bien se intercambian la interpretación de los unos y ceros de la máscara de cruce. Esta técnica es una generalización del esquema de multipunto, en donde el número de puntos de corte M se elige de manera aleatoria para cada reproducción.

3.2.2.4. CRUZA DE CICLO.

Tomamos el primer gen del genoma del padre, poniéndolo en la primera posición del hijo, y el primer gen del genoma de la madre, poniéndolo dentro del genoma del hijo en la posición que ocupe en el genoma del padre. El fenotipo que está en la posición que ocupa el gen del genoma del padre igual al primer gen del genoma de la madre se va a colocar en la

posición que ocupe en el genoma del padre, y así hasta rellenar el genoma del hijo. Es una buena idea que, tanto la codificación como la técnica de cruce, se hagan de manera que las características buenas se hereden; o, al menos, no sea mucho peor que el peor de los padres. En problemas en los que, por ejemplo, la adaptación es función de los pares de genes colaterales, el resultante del cruce uniforme tiene una adaptación completamente aleatoria.

3.2.3. OPERADOR DE MUTACIÓN.

El operador de la mutación es considerado como básico ya que proporciona un pequeño elemento de aleatoriedad en la vecindad de individuos de la población. Si bien se admite que el operador de cruza es el responsable de efectuar la búsqueda a lo largo del espacio de posibles soluciones, el objetivo del operador de mutación es producir nuevas soluciones a partir de la modificación de un cierto número de genes de una solución existente, con la intención de fomentar una variación dentro de la población.

Existen muy diversas formas de realizar la mutación, desde la más sencilla (Puntual), donde cada gen muta aleatoriamente con independencia del resto de genes, hasta configuraciones más complejas donde se tienen en cuenta la estructura del problema y la relación entre los distintos genes.

En la naturaleza se dan cambios en las especies que no se deben a los genes heredados de sus antepasados, sino a las mutaciones que se producen por azar en su cadena genética y pueden aportar valores positivos o negativos en la adaptación de la especie a su entorno. Esto mismo se realiza en los algoritmos genéticos. Se elige al azar cuántos y cuáles de los individuos van a mutar (un individuo no puede mutar más de una vez), se elige que gen de su cadena mutará y se cambia ese gen por su complementado, como se ve a continuación:

1|0|101 => 1|1|101

Ejemplo 3.2. : mutación

Una vez realizadas las mutaciones, ya tendremos la nueva generación, que, previsiblemente tendrá mejores características de adaptación al problema que la anterior, es decir, mejores funciones de *fitness*. Schaffer concluye en su trabajo que determinar el valor óptimo con respecto de la probabilidad de mutación es mucho más crucial que el relativo a la probabilidad de cruza.

Si bien en la mayoría de las implementaciones de los algoritmos genéticos se asume que tanto la probabilidad de cruza como la de mutación permanecen constantes, algunos autores han obtenido mejores resultados experimentales modificando la probabilidad de mutación a medida que aumenta el número de iteraciones.

Se deben realizar numerosos ciclos de selección, cruce y mutación para mejorar suficientemente la población de individuos (Figura 3.6.). El algoritmo genético debe finalizar cuando considere que la mejor solución encontrada hasta ese momento es bastante buena o que no va a poder encontrar una mejor. Para eso se define una condición de parada.

La condición de parada más sencilla es hacer que el algoritmo pare siempre después de haber hecho un número de ciclos predeterminado. Otra posible condición de parada es que la función de *fitness* alcance un valor determinado que se considere que es suficientemente bueno para ese problema. La condición que se utilizará en este trabajo como condición de parada es una mezcla de ambas.

Se guardará el individuo que tenga el valor de *fitness* más alto de la generación inicial. A partir de ahí, si ese valor no es superado en veinte ciclos, el algoritmo parará. Siempre que el valor de *fitness* del individuo sea superado, se guardará el nuevo mejor individuo y se volverá a contar hasta veinte ciclos del algoritmo. La solución al problema será el individuo que esté guardado en el momento que pare el algoritmo.

A continuación se muestra el diagrama de flujo del algoritmo:

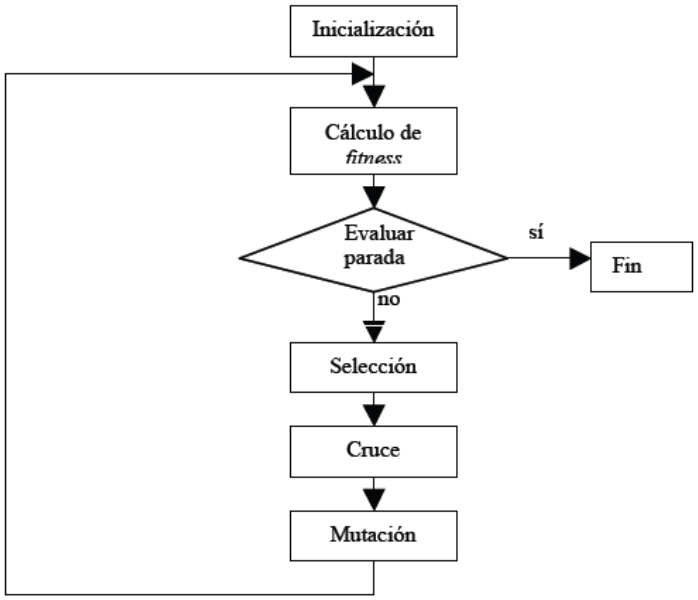

Figura 3.6. Diagrama de flujo del algoritmo

IV. IMPLEMENTACIÓN DE LOS ALGORITMOS

Para la implementación hemos hecho uso de un algoritmo genético simple, así como para la aplicación de clustering, utilizaremos la metodología de racimo por semillas y una base de datos que contiene valores numéricos que corresponden a la captación durante un año de un determinado gas en la atmosfera en diferentes puntos de la Zona Metropolitana del Valle de México. Cabe destacar que estos valores fueron generados en tiempo real por un sistema que nos permitió evaluar las partículas en el ambiente y transformarlas en valores numéricos, es así como se determinó que por la inmensa cantidad de datos que se almacenados en nuestra base de datos, se consideró como herramienta de trabajo MATLAB, que permitirá tener un manejo sobre los datos de una manera eficaz y eficiente, así como también del procesamiento de los datos que se tiene que realizar, debido a la magnitud de nuestros datos que facilitara los cálculos en un periodo más corto de tiempo.

A partir de observar la numerosa cantidad de datos que tenemos, se desarrolló una interfaz sencilla (Figura 4.1.) – en el entendido, que nos referimos al manejo de la misma - que permitirá manipular algunos valores de importancia (la cruza y la mutación) antes de realizar cualquier procesamiento con los datos. Este sistema contará con algunos menús que facilitarán al usuario tener un manejo sencillo y entendible para su correspondiente estudio. El usuario tendrá en sus manos la posibilidad de realizar hacer los comparativos que considere pertinentes con respecto a los valores que se encuentran almacenados en los data sets y su respectiva aplicación del algoritmo genético, dentro de nuestra interfaz hemos colocado las siguientes características:

Menú LOAD DB, es el encargado de seleccionar el archivo donde se encuentra almacenados los datos para su evaluación, cabe destacar que nuestros archivos se encuentran debidamente guardados en documentos de Excel, lo cual permite hacer uso de algunas funciones en MATLAB para la manipulación de los datos.

Menú % MUTACIÓN, es un menú de despliegue que permite realizar una selección que indicará el porcentaje de mutación que tendrán los datos, la eficiencia de la mutación está

en dependencia del problema y en nuestro caso particular podremos hacer uso de los valores de 0.0 hasta 1.0 en nuestra interfaz.

Menú No. GENERACIONES, se determina en cuantas ocasiones consideramos que es suficiente que se realice la aplicación de nuestro algoritmo genético sobre nuestros valores, esto ayudará para obtener la condición de paro de nuestro problema.

Los menús ALFA y BETA nos permiten colocar valores que tendrán efecto en el numero de semillas, sin que se defina una cantidad que sea idónea para un mejor resultado se recomienda que los valores que sean tomados en cuenta sean 0.5 y 1.0 respectivamente.

Menú No. ESTACIÓN, este menú permitirá hacer la selección de una estación que contenga información sobre la presencia del gas en la atmosfera, las estaciones serán seleccionadas mediante un valor numérico, en el caso de que el valor que se asignara no existiera o no contenga información se dará a conocer en la consola del programa. Cabe destacar que todos los gases difieren del número de estaciones en las cuales fueron captadas.

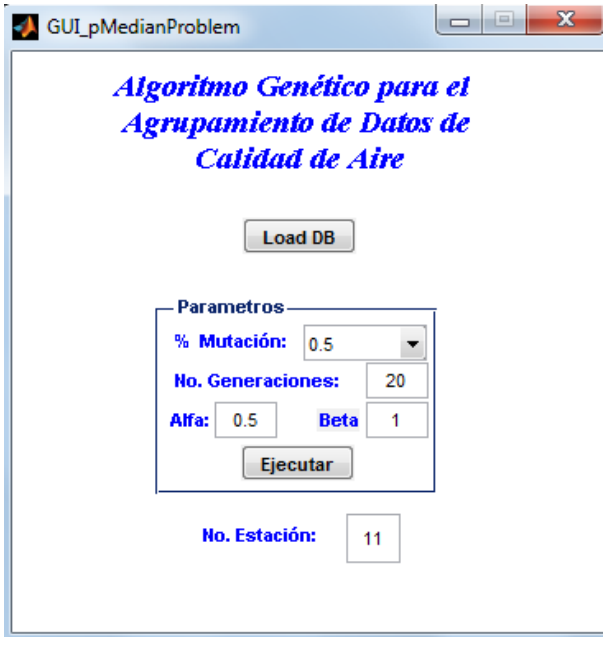

Figura 4.1. Interfaz.

Botón EJECUTAR, finalmente esta opción será la encargada de dar la instrucción para dar inicio a las operaciones que se programaron tomando las consideraciones que el usuario haya considerado como idóneas en la interfaz para observar el comportamiento de los datos que posteriormente serán mostradas en las diferentes gráficas para su análisis.

Para la implementación se hace mención que se utiliza un proceso de clustering sobre los datos al evaluar y el uso del algoritmo genético conocido como agrupamiento por racimo de semillas ya que sea considerado en algunos otros artículos de investigación como el más optimo debido a sus características y su desempeño al evaluar datos sobre problemas de gran magnitud, tales consideraciones y su justificación ya ha sido explicada en este mismo trabajo.

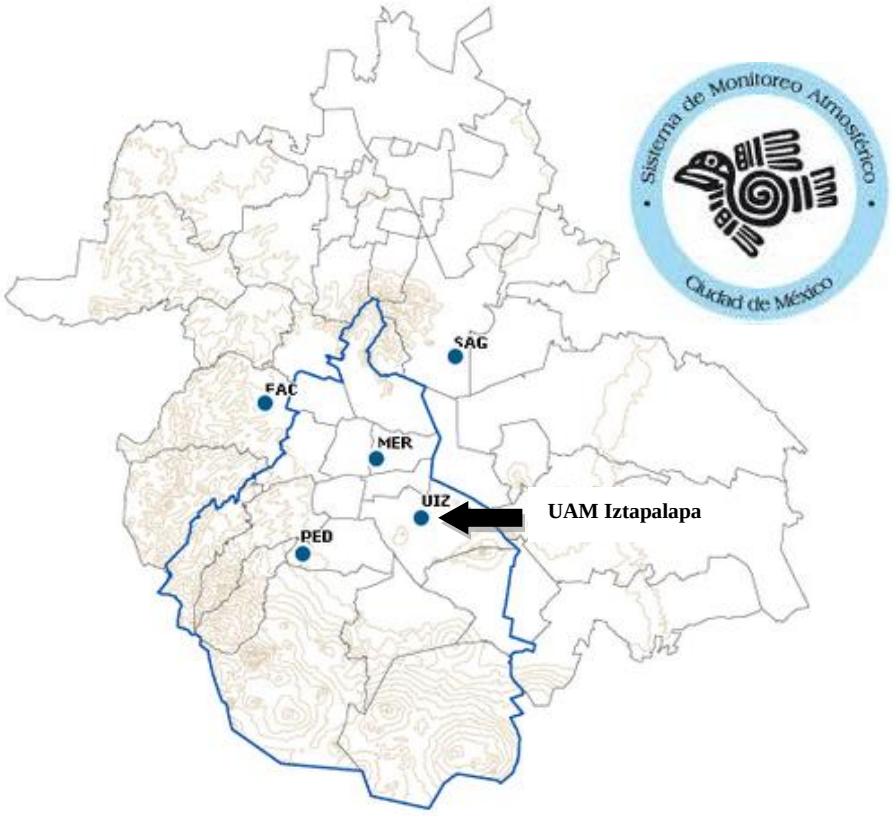

Figura 4.2. Mapa de la ubicación en MZVM de la estación a evaluar del SIMAT.

Es así como en los resultados que se muestran de manera gráfica después de haber realizado las operaciones sobres los datos. En este trabajo se observará el comportamiento de la estación UAM Iztapalapa (UIZ) (Figura 4.2.) que cuenta con información de la existencia de Monóxido de Carbono de las mediciones realizadas por la Red de Automática de Monitoreo Atmosférico (RAMA), que pertenece al Sistema de Monitoreo Atmosférico (SIMAT).

El algoritmo genético tiene como objetivo obtener un conjunto de semillas de tamaño par (o en algunos casos muy remotos el conjunto de las semillas es igual a 1), que inicialmente determinan el *fitness* de las semillas que se van a tomar de entre la población, nótese que las semillas generadas dependerán de los valores asignados por "Alfa" y "Beta".

Se coloca en un arreglo todos los datos de nuestra estación (con un total de 8760 elementos), que serán divididos para su evaluación en bloques de 24 elementos debido a que nuestros resultados serán mostrados a partir de el número de horas en un día con lo que obtendremos 365 bloques que corresponden a los días en una año. Los datos serán normalizados de manera tal que se garantice que nuestros datos tengan un valor estrictamente numérico. Se crea una matriz donde los elementos son verificados en base a su posición colocando un 1, de aquellos elementos que se encuentren debidamente ubicados en su posición correcta (es decir, en suposición i, j donde i = j) se generan las semillas utilizadas para mejorar nuestra población y así dar inicio a nuestro algoritmo genético y aquellos elementos que se encuentren fuera de la posición correcta serán destinados hacer los elementos a mejorar (son los defectuosos).

En el proceso del algoritmo genético existe un *fitness* o aptitud para cada individuo de cada generación, que será el encargado de mejorar en n – generaciones aquellos elementos de la población que no se encuentren en su posición correcta, el valor de n corresponde al número de generaciones que se considere necesario para interactuar en el algoritmo, sin que se garantice que al término tenga la obligación de mejorar a toda la clase defectuosa. Este proceso continuara para cada bloque de 24 elementos (en 365 ocasiones) y para cada bloque se genera un arreglo que ira agrupando a la población por sus características es decir el primer vector corresponderá a los datos que pertenezcan a los valores de la hora uno, el

segundo vector tendrá los datos que corresponda a los valores de la hora dos y así sucesivamente hasta obtener veinticuatro arreglos. Posteriormente se presentarán las gráficas correspondientes, en primer lugar es una gráfica representativa de la variación promedio por horas del gas que se evaluó (Figura 5.5. y Figura 5.6.), la segunda gráfica nos mostrara la variación promedio del gas en periodos de cuatro horas (Figura 5.7. y 5.8.) y la última gráfica presenta la frecuencia de clusters utilizados durante el año para el agrupamiento de los datos (Figura 5.9. y 5.10.).

V. RESULTADOS Y APLICACIONES

Para ejemplificar la teoría descrita en el trabajo se realizaron pruebas con documentación acerca de Monóxido de Carbono del año 1995 y 2010 respectivamente, para hacer una comparativa del comportamiento de nuestro algoritmo genético en archivos que contienen información de 15 años de distancia. La concentración promedio anual de Monóxido de Carbono (CO) en el Distrito Federal se muestra en la figura 5.1.

El monóxido de carbono es un gas incoloro, inodoro e insípido, poco soluble en agua. Está formado por un átomo de carbono ligado covalentemente a un átomo de oxígeno. Es poco soluble en agua y su densidad es ligeramente menor que la del aire. El monóxido de carbono es un compuesto altamente tóxico, sin embargo juega un papel importante en la síntesis y producción de una gran cantidad de productos. El origen del CO es diverso, entre las fuentes naturales que lo producen se encuentran la quema de biomasa y la oxidación de compuestos orgánicos como el isopreno y el metano.

El CO es altamente tóxico para los seres humanos y otras formas de vida aeróbicas, inhalado en pequeñas cantidades puede producir hipoxia, daño neurológico y posiblemente la muerte (Tabla 5.1.). Aún en concentraciones pequeñas como de 400 ppm en el aire, el CO puede ser fatal. Una característica peligrosa está relacionada a que carece de olor, lo cual da lugar a que no sea detectado por el olfato del ser humano. Los primeros síntomas del envenenamiento por CO pueden ser mareo y dolor de cabeza, seguidos de inconsciencia, falla respiratoria e incluso la muerte.

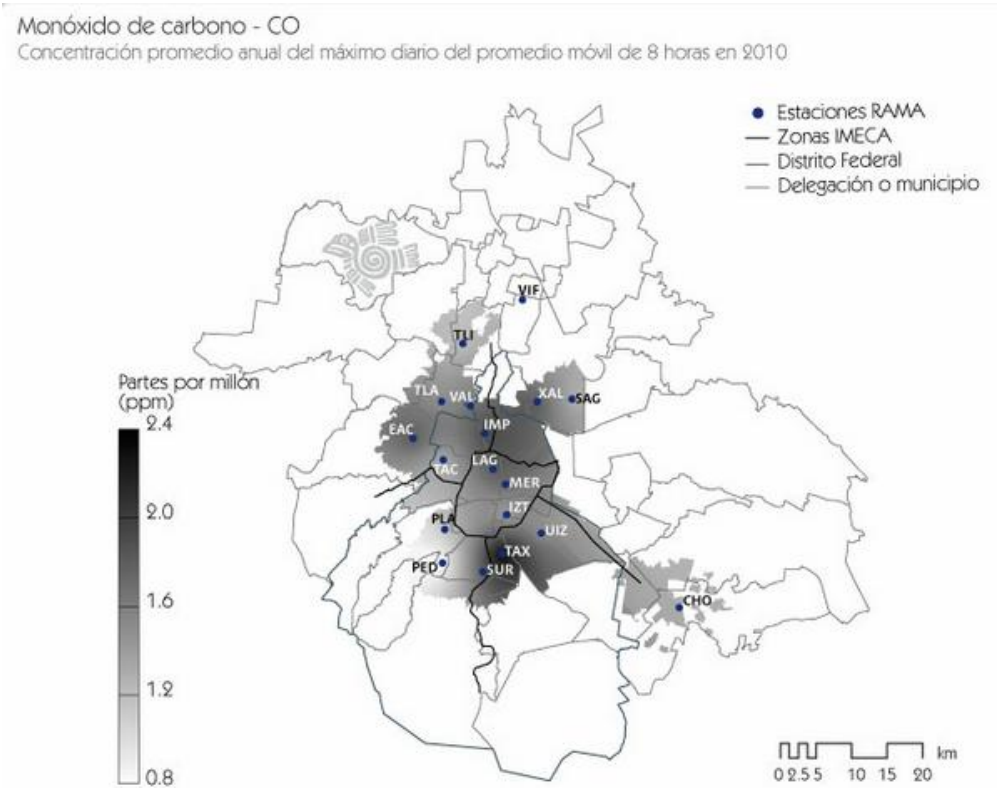

Figura 5.1. Mapa sobre la presencia de Monóxido de Carbono en MZVM.

Tabla5.1. Daños a la salud provocadas por Monóxido de Carbono.

Concentración de CO	**Tiempo de exposición**	**Daños a la salud**
35 ppm	6-8 hrs	Dolor de cabeza y mareo
100 ppm	1-2 hrs	Dolor de cabeza leve
200 ppm	1-2 hrs	Dolor de cabeza leve
400 ppm	1-2 hrs	Dolor frontal de cabeza
800 ppm	45 minutos	Mareo, nauseas y convulsiones
1,600 ppm	20 minutos	Dolor de cabeza, mareo y nausea

En México la Norma Oficial Mexicana NOM-021-SSA1-1993 establece un límite máximo permisible de 11.0 partículas por millón (PPM) en un promedio móvil de 8 horas, lo cual no debe excederse más de una vez al año (Figura 5.2.). La Norma Oficial Mexicana NOM-034-SEMARNAT-1993, emitida el 18 de Octubre de 1993, establece los métodos de medición para determinar la concentración de monóxido de carbono en el medio ambiente y los procedimientos para la calibración de los equipos de medición (Figura 5.3.).

De acuerdo con los datos del monitoreo atmosférico, desde 1991 la concentración de monóxido de carbono en el aire ambiente ha presentado una tendencia a la baja. La renovación del parque vehicular, la instalación de sistemas de inyección electrónica y la incorporación de los convertidores catalíticos en los vehículos que circulan en la Ciudad de México son las medidas qua han impactado de manera significativa en la disminución de este contaminante (Figura 5.4.).

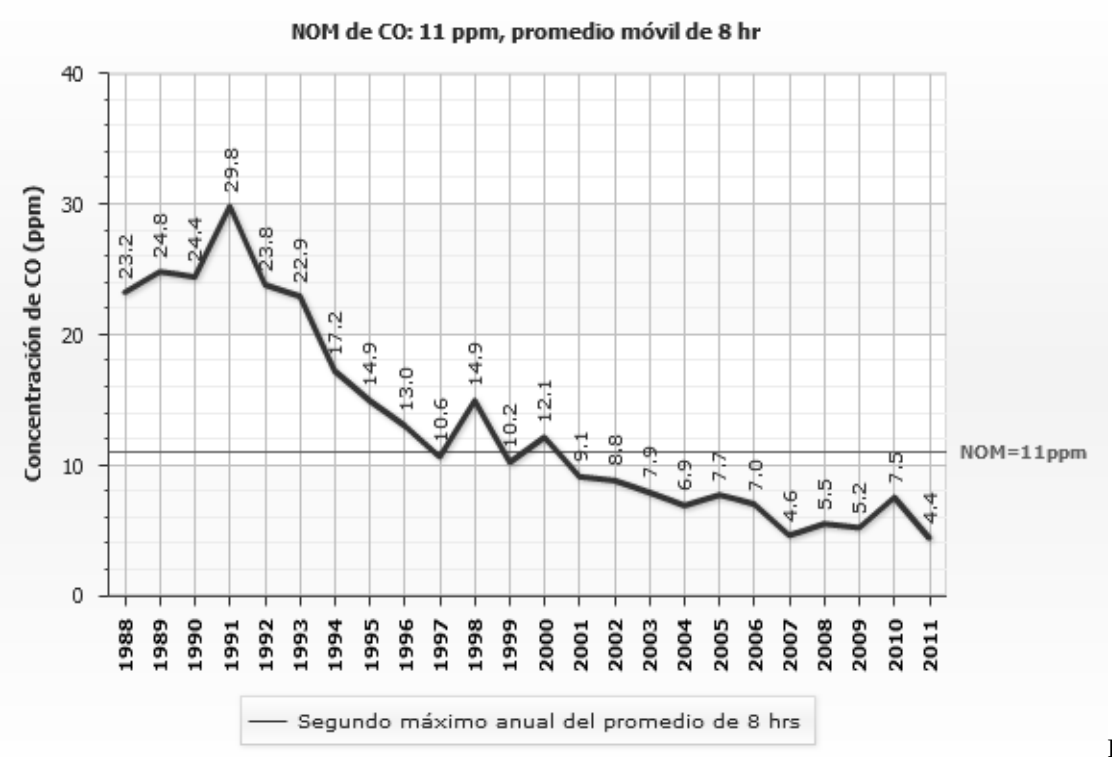

Figura 5.2. Concentración promedio móvil de 8horas de CO.

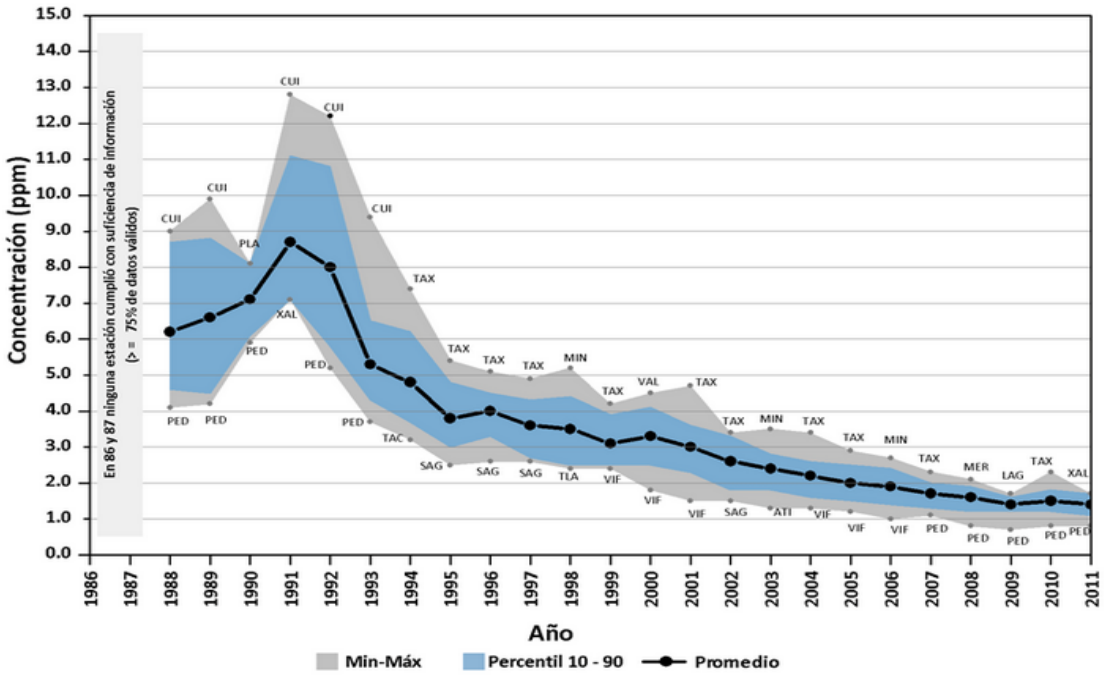

Figura 5.3. Concentración de PPM de CO.

Para la evaluación del trabajo de investigación los documentos contienen 8760 datos por estación cada uno, que fueron captados por los sistemas correspondientes y clasificados en días y estos divididos en periodos de 24 horas durante un año, haremos las evaluaciones correspondientes con los datos obtenidos del año de 1995 (Tabla 5.3.). Posteriormente mostraremos el comportamiento de los datos obtenidos en el año de 2010 (Tabla 5.4.), teniendo estos resultados haremos las comparativas de sus graficas y su interpretación para así mostrar el funcionamiento resultante de nuestro algoritmo genético y su comportamiento sobre los datos. Los parámetros que utilizaremos en la interfaz (Tabla 5.2.) para nuestra comparativa son los siguientes:

Tabla5.2. Parámetros establecidos para evaluar en la interfaz.

Estación	% Mutación	No. Generaciones	ALFA	BETA
UAM Iztapalapa (UIZ)	0.5 %	20	0.5	1.0

Tomando en cuenta las anteriores consideraciones los resultados que se obtuvieron fueron los siguientes:

Tabla5.3. Datos del años de 1995, arrojados por el algoritmo después de su evaluación.

UAM Iztapalapa (UIZ) 1995					
Ejecución	**Elementos**	**Semillas**	**Clusters**	**Clusters Prom. – Dia**	**Tiempo de ejecución (seg)**
1	8760	16	77	3.9035088	21.58685
5	8760	10	66	3.7022901	30.96810
10	8760	2	83	4.553719	46.54003
20	8760	2	77	4.2347826	39.81307

Tabla5.4. Datos del años de 2010, arrojados por el algoritmo después de su evaluación.

UAM Iztapalapa (UIZ) 2010					
Ejecución	**Elementos**	**Semillas**	**Clusters**	**Clusters Prom. – Dia**	**Tiempo de ejecución (seg)**
1	8760	12	81	4.4869565	20.46320
5	8760	18	97	3.8695652	22.63772
10	8760	1	71	3.952381	45.96687
20	8760	6	73	5.82191781	21.01314

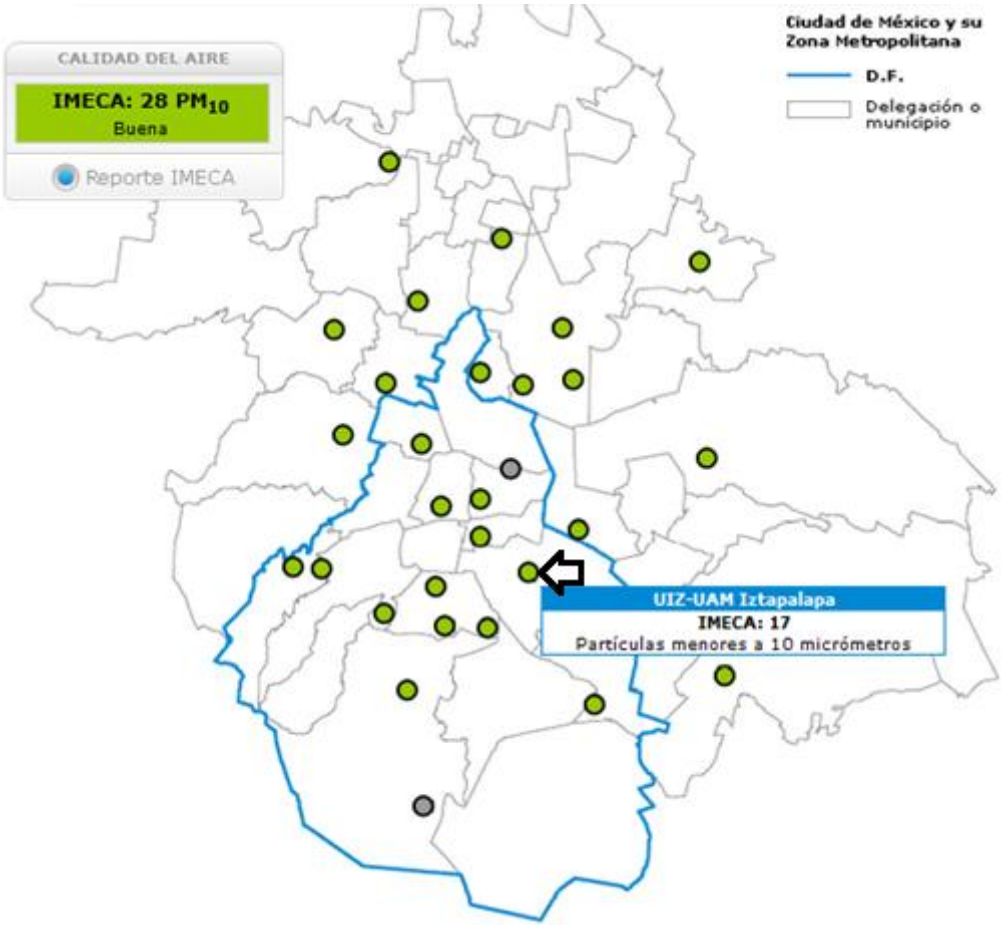

Figura 5.4. Mapa de las estaciones del Valle de México.

Como se puede observar en las tablas anteriormente mostradas (Tabla 5.3. y Tabla 5.4.) se realizarán el mismo número de corridas y arrojaron como resultados una serie de comportamientos diversos, en la ejecución 1 se observa que en los dos procesos se obtuvieron un gran número de semillas generadas y se observó que nuestros elementos tuvieron un número igualmente grande de clusters, aquí la diferencia se da en el promedio - día que tuvieron las ejecuciones, es decir, mientras que la tabla 5.3. del año de 1995 muestra un promedio de clusters – día cercano a los 4, la tabla 5.4. del 2010 generó un promedio cercano a los 4.5 clusters, es decir, tuvo un mayor uso el algoritmo para el proceso de clustering o agrupamiento.

Aunque lo que respecta a tiempo de ejecución se observa que fue muy similar debido a que el algoritmo de racimo por semillas destaca así como ya se ha justificado en este documento por una mayor eficiencia y eficacia en ejecución. De igual forma podemos hacer el comparativo en la ejecución número 20 en este caso en particular se observa que hay una mayor cantidad de semillas generadas en el año 2010 con respecto de 1995 sin embargo existe una mayor cantidad de clusters en el año 1995, aun así, el promedio de clusters – día es mucho mayor en el año de 2010 al igual que el tiempo de respuesta de nuestro algoritmo genético fue mejor.

En las gráficas que se presentan a continuación (Figura 5.5. y Figura 5.6.) se puede observar la variación promedio por hora durante un año de los puntos IMECA, que es la medición que nos permite evaluar con base a estos parámetros la calidad de aire durante un año (Tabla 5.5.).

Tabla5.5. Interpretación del IMECA.

INTERPRETACIÓN DEL IMECA	
IMECA	**CONDICIÓN**
0-50	Buena
51-100	Regular
101-150	Mala
151-200	Muy mala
>200	Extremadamente Mala
M/F	Mantenimiento o falla del equipo

Por ejemplo, se puede observar que en la “hora 1” de la gráfica de 1995 muestra una condición buena en base a la tabla de interpretación del IMECA que da una valor de 10, es decir, la calidad del aire permanece estable, sin embargo si comparamos la misma “hora 1” en la gráfica del año 2010 que se encuentra de igual manera en una condición “Buena” pero

tiene un valor de 8,es decir, que la calidad de aire ha mejorado en 15 años. Cabe destacar que no en todas las 24 horas se encuentra una mejoría en la calidad de aire esto es que no siempre se encuentran los mismos factores de alteración en su medición.

Aun así, en la mayoría de los casos si se presenta una mejoría en la calidad de aire con respecto al Monóxido de Carbono por lo tanto se concluye que en la estación que corresponde a UAM Iztapalapa (UIZ) 2010 no se ha detectado algún peligro considerable para la población que habita en los alrededores donde se encuentra situada la estación.

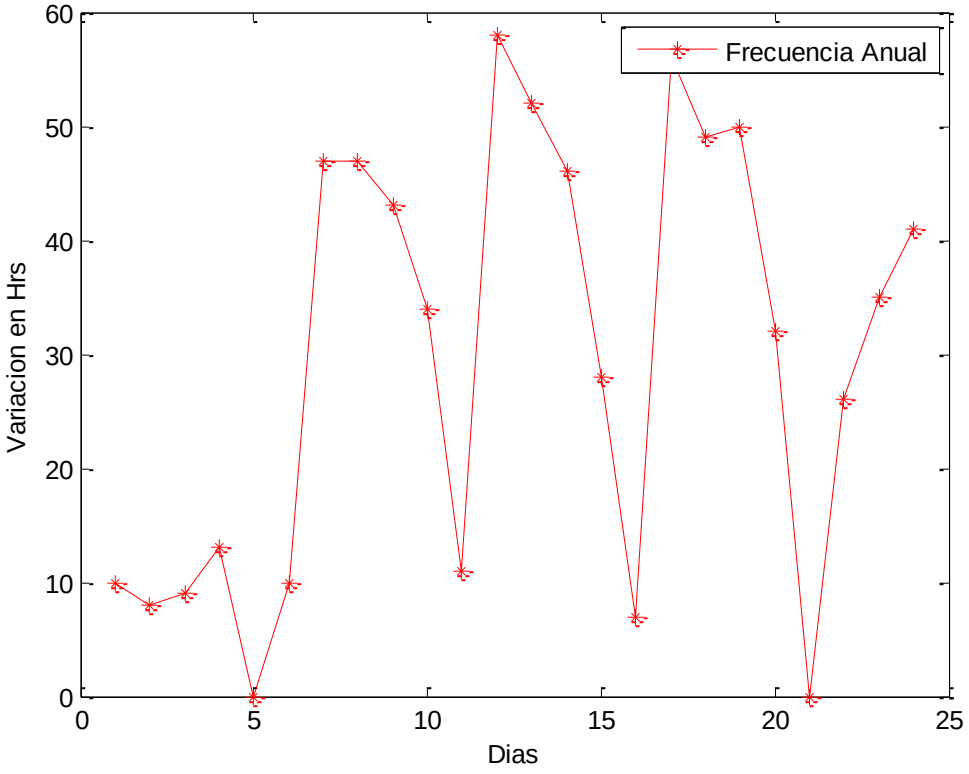

Figura 5.5. Variación promedio por hora IMECA, Año 1995.

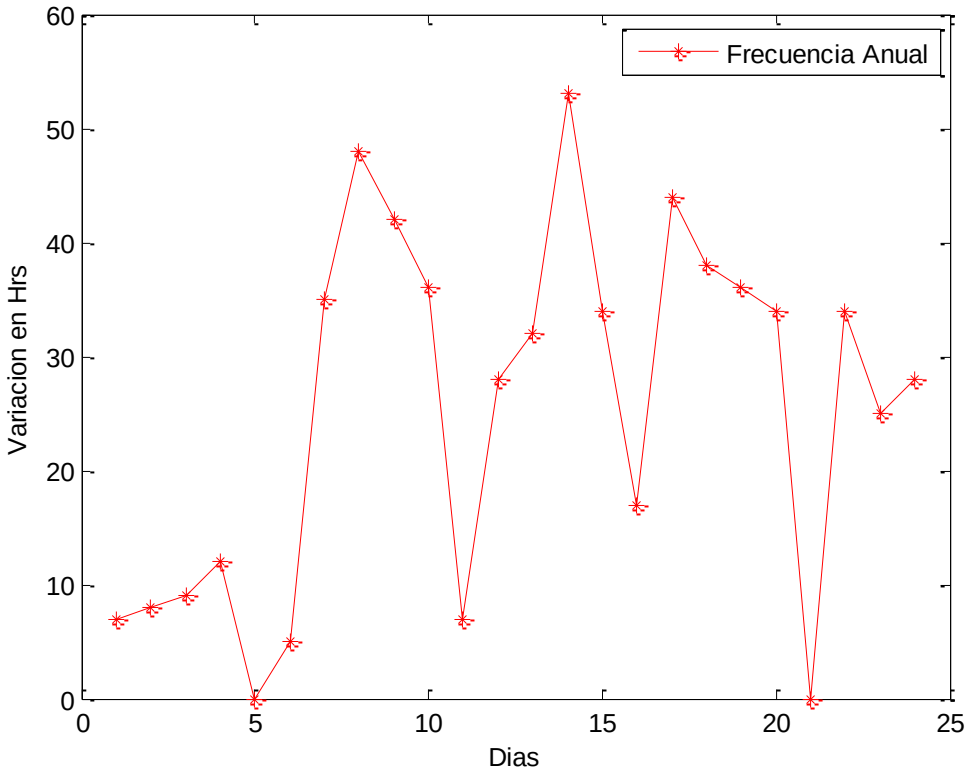

Figura 5.6. Variación promedio por hora IMECA, Año 2010.

En las siguientes gráficas (Figura 5.7. y Figura 5.8.) podremos observar la variación promedio en periodos de 4 horas de la presencia de Monóxido de Carbono CO que sufrió en un año. En la gráfica que corresponde al año de 1995 (Figura 5.7.) se ve claramente que en las primeras cuatro horas del año su variación fue la menor y alcanzó un 0.06%, mientras que el conjunto que comprende de las 16:00 horas hasta las 20:00 horas se observa la mayor variación de presencia de Monóxido de Carbono marcado en 0.27%, dicho de otra forma fueron las horas que mayor riesgo presentaron para la población en 1995. La diferencia de variación de porcentaje fue de 0.21%. En la gráfica del año de 2010 (Figura 5.8.) sus valores de variación correspondieron a la menor de 0.07% y la mayor se presentó en 0.25%, es decir que la diferencia de la presencia de CO fue menor a lo presentado en el año de 1995 por lo que nos muestra una mejora en nuestra calidad de aire en la estación evaluada.

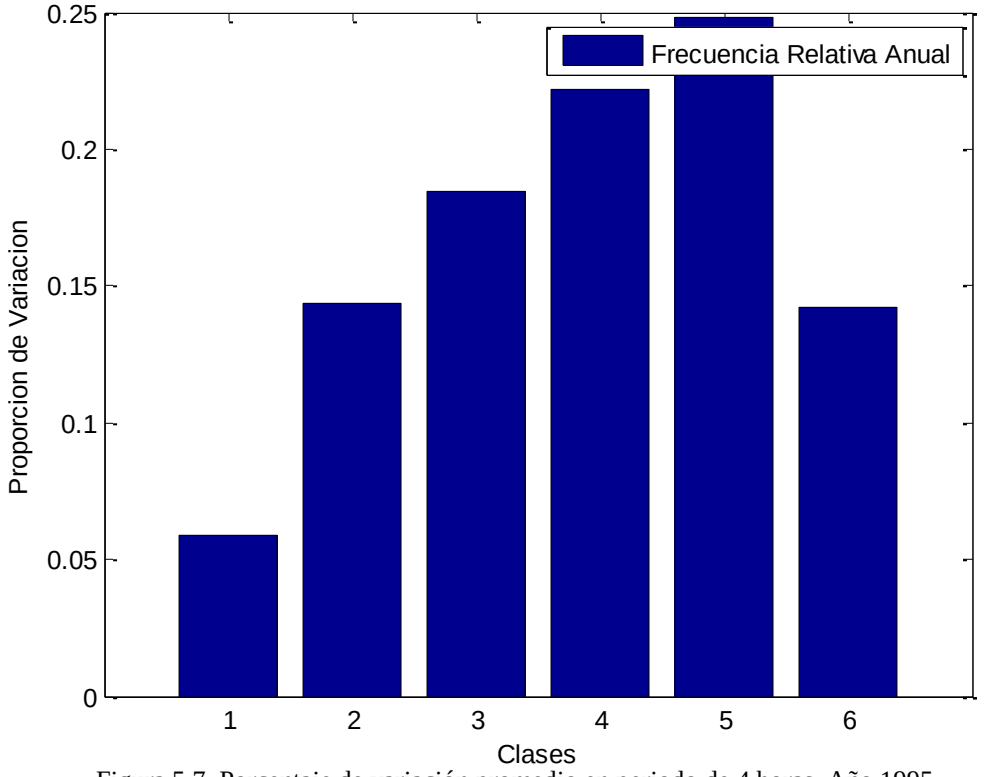

Figura 5.7. Porcentaje de variación promedio en periodo de 4 horas, Año 1995.

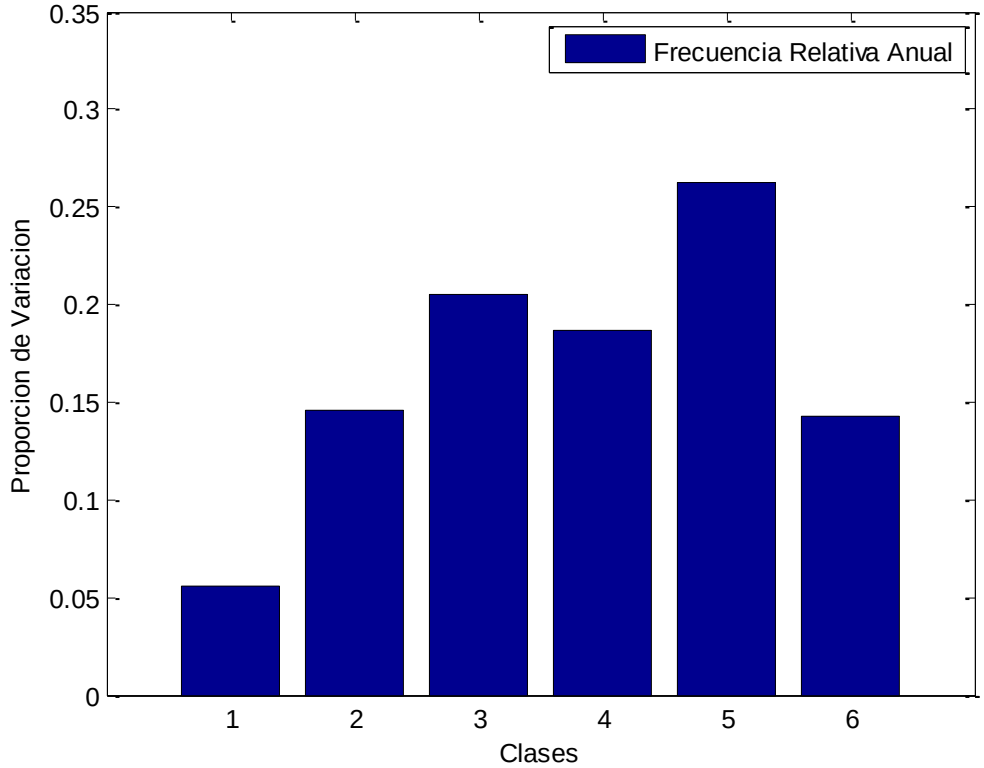

Figura 5.8. Porcentaje de variación promedio en periodo de 4 horas, Año 2010.

En los siguientes dos gráficas (Figura 5.9. y Figura 5.10.) se muestra el número de clusters que tuvieron frecuencia anual que está dada en días, en el año de 1995 presenta 29 días con 3 clusters que fueron los que mayormente se utilizaron durante el año y hubo 11 clusters que solo se presentaron en 2 días a lo largo del año. Cada gráfica presenta el número total de clusters utilizados a lo largo del año y en la tabla (Tabla 5.3. y Tabla 5.4.) se encuentran los valores obtenidos en las diferentes ejecuciones y el promedio anual de clusters que fueron utilizados.

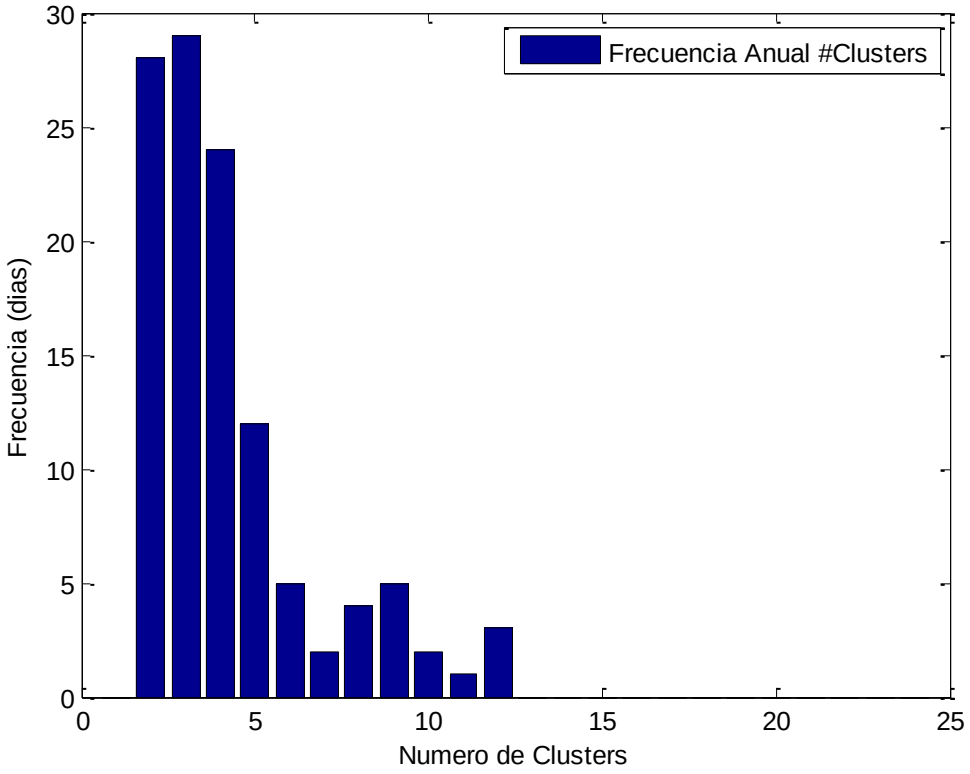

Figura 5.9. Frecuencia anual de clusters, Año 1995.

En el año de 2010 se presentaron 27 días con dos y tres clusters utilizados para el agrupamiento de los datos siendo estos los que mayor frecuencia en días presentaron y los clusters que menor uso de días obtuvieron son los de 12 y 16 clusters.

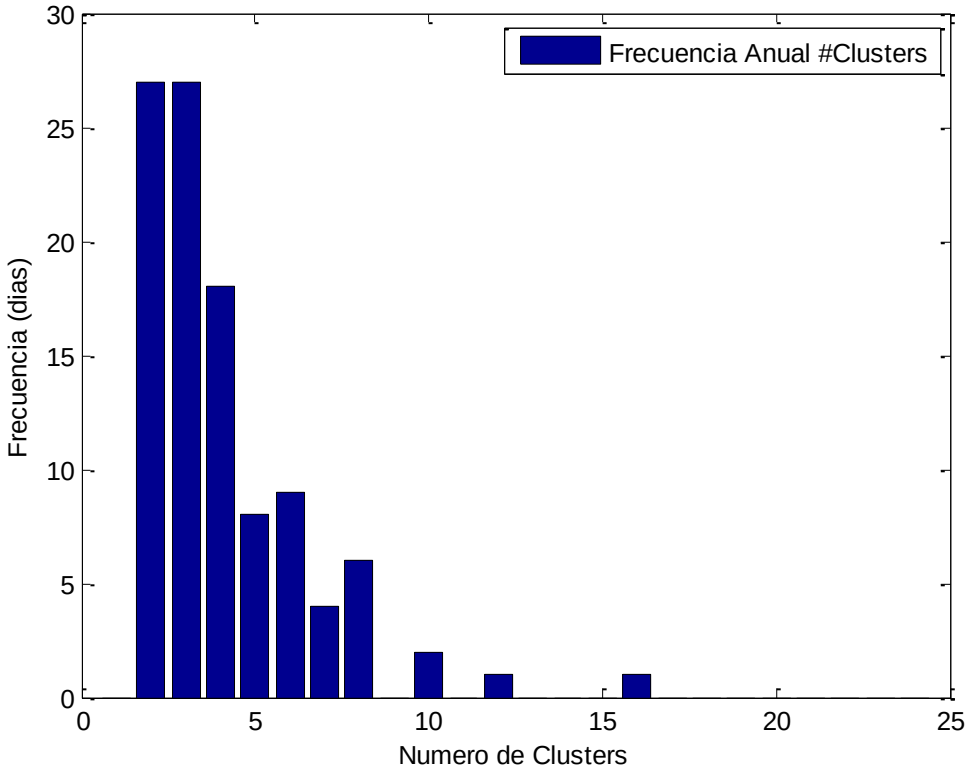

Figura 5.10. Frecuencia anual de clusters, Año 2010.

La evaluación de factores que intervengan en la afectación de la calidad de aire ya sea por crecimiento de la población, efectos naturales o algunos otros se han convertido en indispensables para tratar de mantener ciertos niveles que nos permitan controlar las afectación que pudieran impactar en la población. Hoy en día la contaminación por Monóxido de Carbono no representa una problemática incluso se ha notado una disminución considerable en los últimos años de los promedio anuales de concentraciones diarias en la mayorías de las ciudades del país. Los algoritmos genéticos deben seguir una línea de estudia para perfeccionar técnicas que nos permitan realizar evaluaciones más directas que nos permitan resolver o predecir en periodos de tiempo posibles emisiones que se pudieron presentar de las diferentes monitoreos sobre la calidad del aire.

VI. CONCLUSIONES Y TRABAJO A FUTURO

Los algoritmos genéticos permiten utilizar su funcionamiento para evaluar problemas de cualquier área de investigación, o incluso una propuesta de optimización en algún área en específico. El mundo para elaborar los algoritmos genéticos es inmenso, por mencionar las áreas en las que se ha desarrollado o implementado este tipo de técnicas y modelos se encuentran en la medicina, el área automotriz, la economía, la administración, en la ecología, sistemas sociales, aprendizaje automático, etc.

Aunque, esta lista no es, en modo alguno, exhaustiva, sí transmite la idea de la variedad de aplicaciones que tienen los Algoritmos Genéticos. Gracias al éxito en éstas y otras áreas, los Algoritmos Genéticos han llegado a ser un campo puntero en la investigación actual.

Además requieren en la mayoría de los casos poca información sobre el espacio de búsqueda, ya que están diseñados para trabajar sobre conjuntos de soluciones, es decir, buscan una solución por aproximación de la población y puede llegar hacer métodos generalizados.

Con este trabajo se muestra la eficiencia y la eficacia de los algoritmos genéticos y el proceso de agrupamiento de datos (clustering) para resolver problemas complejos, además no se requiere de mucho tiempo de ejecución (dependiendo el tamaño del problema). El comportamiento del algoritmo genético permitió obtener en la gran mayoría de los casos una respuesta satisfactoria para el proceso de clustering, ya que permitía tener un agrupamiento de los datos relativamente similar basado en sus características (su valor numérico).

Debemos dejar claro que los algoritmos genéticos nos permitirán darnos una respuesta óptima mientras mayor sea el número de ejecuciones realizadas al problema. Otro factor importante a considerar fue las herramientas que nos ofrece MATLAB para problemas de optimización, y que en nuestro caso nos ayudó para realizar las presentaciones gráficas de nuestro trabajo.

Sería interesante observar el comportamiento donde tuviéramos la oportunidad de elegir entre algunos operadores de mutación, cruza y selección, que afectan directamente a los datos. O realizar el proceso de forma paralela con dos algoritmos genéticos podría inducirnos a mejores resultados, bastaría intentar realizar las operaciones con dos o más algoritmos genéticos que nos permitieran obtener una solución óptima promedio, así como introducir la mutación como una parte "natural" del proceso de evolución esto con la finalidad de reducir o eliminar clusters que se consideren innecesarios después de ser evaluados los algoritmos genéticos en algunas generaciones. Por otra parte podría implementarse hacer la evaluación de clusters mediante metodologías que nos permitan elegir cierto tipos de clusters que ayuden a la selección de los individuos.

VII. BIBLIOGRAFÍA.

[1] John H. Holland, "Adaptation in Natural and Artificial System", University of Michigan, 1992.

[2] Secretaria de Medio Ambiente y Recursos Naturales, "Informe de la Situacion del Medio Ambiente en Mexico", 2008, http://www.semarnat.gob.mx.

[3] Ulrich Bodenhofer,"Genetic Algorithms", 2003-2004.

[4] Abdelmalik Moujahid, et da, "Algoritmos Geneticos", Departamento de Ciencias de la Computación e Inteligencia Artificial Universidad del Pais Vasco-Euskal Herriko Unibertsitatea.

[5] Carlos A. Coello Coello, "Algoritmos Genéticos y sus Aplicaciones".

[6] Santana Quintero, Luis Vicente, et ad, "Una Introducción a la Computación Evolutiva y Algunas de sus Aplicaciones en Economía y Finanzas", Departamento de Computación. CINVESTAV-IPN (Grupo de Computación Evolutiva), 2006.

[7] Matias Ison, Jacobo Sitt, "Algoritmos Genéticos: Aplicación en MATLAB", 2005.

[8] Ana Clara Vélez Torres, "Un Modelo Genético para Minería de Datos", Universidad Nacional de Colombia Facultad de Minas Maestría Ingeniería de Sistemas, 2001.

[9] José Manuel Molina López, "Técnicas de Análisis de Datos", Universidad Carlos III de Madrid, 2006.

[10] Andrés Bravo Méndez, "Algoritmos Genéticos".

[11] L. Recalde, "Esquemas algorítmicos - Algoritmos Genéticos".

[12] José Luis Lezama, "Los grandes problemas de México", Colegio de México, 2010.

[13] Marcos Gestal Pose, "Introducción a los Algoritmos Genéticos", Departamento de Tecnologías de la Información y las Comunicaciones Universidad de Coruña.

[14] Gregorio Toscano Pulido, "Estrategias Evolutivas".

[15] "Estrategias Evolutivas", University of Nottingham, 2004.

[16] Piedad Tolmos Rodríguez-Piñero, "Introducción a los algoritmos genéticos y sus aplicaciones".

[17] Rosa María Yáñez Gómez, "Introducción a las tecnologías de clustering en GNU/Linux".

[18] Ruth A. Etze, "Contaminación del aire".

[19] A.K. JAIN et ad, "Data Clustering: A Review", Michigan State University.

[20] Manuel Terrádez Gurrea, "ANÁLISIS DE CONGLOMERADOS", www.uoc.edu.

[21] Minerva Catalán-Vázquez et ad, "Percepción de riesgo a la salud por contaminación del aire en adolescentes de la Ciudad de México".

[22] NATYHELEM GIL LONDOÑO, "ALGORITMOS GENETICOS", Universidad Nacional de Colombia Escuela de Estadística Sede Medellín, 2006.