Benemérita
Universidad Autónoma de Puebla

Facultad de Ciencias de la Computación

Descubriendo conocimiento en bases de datos de cambio climático con técnicas de cómputo suave

TESIS
PROFESIONAL

Que para obtener el grado de:

Maestro en Ciencias de la Computación

Presenta:

Jaime Enrique Reyes Salazar

Asesor:

Dr. Abraham Sánchez López

Puebla, Pue. Otoño 2012

Dedicado a mi familia,
que me ha apoyado en todo momento para concluir mi maestría.
Gracias mamá (Eva Salazar L.) y papá (Jaime Reyes A.).
Gracias a mis hermanitas Shary y Balvi.

Agradecimientos

Quiero iniciar por agradecer a mis padres: Eva Salazar Lucero y Jaime Reyes Ama, que me han brindado la oportunidad de vivir, de disfrutar al máximo de la vida, de reir, de soñar, y por supuesto de estudiar. Gracias a los dos por su cariño, sus enseñanzas, su tiempo, su paciencia, sus regaños y sobre todo su amor.

A mi asesor el Dr. Abraham Sánchez López, por ser un gran ejemplo a seguir para esforzarme por realizar las cosas.

A mis "manas" Shary y Balvi, pues siempre es divertido pensar en un fin de semana a su lado, jugando, riendo y hasta siendo golpeado por ellas. Las quiero mucho hermanas!.

A mis tios: Norbe, Maricela, Hugo, Beto, Javier y a mis primos: Chooper, Pipo, Genaro, Rosa, Sara, Juan y Baldo. Por brindarme una infancia inolvidable llena de recuerdos bonitos, canciones, poesía y chistes.

A los profesores de la Facultad de Ciencias de la Computación, en especial a la M.C. Yalú Galicia, y al maestro Armando Espindola pues con sus enseñanzas, su paciencia y su tiempo me inspiraron a lograr este trabajo. Así mismo me gustaría agradecer al Consejo Nacional de Ciencia y Tecnología (CONACYT) por la beca con la que se me apoyó para la realización de estos estudios.

A todos mis amigos, en especial a esas personas importantes en mi vida: Amanda, Rox, Amparo, Yarid, Coni, Caro, Popis, Liz, Julissa, Win, Miyo, Chooper, Durango, Rostro, Fortiz, Chiquis e Isaias; gracias por brindarme su amistad.

A los compañeros de los laboratorios de: Educación Continua, Robótica Móvil y Movis Group; así como a mis compañeros de inglés y Frances.

Resumen

El análisis de datos es sumamente importante, pues gracias a este podemos descubrir conocimiento, lo que nos permite comprender el problema con el que estamos tratando, además de brindar soluciones, basadas en la información adquirida durante el proceso de descubrimiento.

El descubrir conocimiento es un área dentro de las Ciencias de la Computación, en donde se usan técnicas de Inteligencia Artificial, Aprendizaje Artificial, Cómputo Suave y Procesamiento del Lenguaje Natural; con la finalidad de procesar los datos de un problema en particular para inferir conocimiento que nos permita detallar características que no son tan evidentes en los datos.

Una técnica para el análisis de características de datos donde no se cuenta con grupos definidos, es el agrupamiento de estos usando una medidia de similaridad. Esto nos permite analizar el comportamiento, así como las características que determinan los grupos.

Sin embargo el procesamiento de grandes cantidades de datos, hace que las técnicas clásicas de agrupamiento se tornen lentas, al agrupar los objetos a ser analizados. Es por ello que es necesario generar algoritmos que combinen estrategías computacionales con la finalidad de encontrar el agrupamiento óptimo.

Una estrategía de optimización son los algoritmos genéticos pues estos se basan en la teoría de la evolución para iterar a través de soluciones al problema y aplicar operadores de cruza y mutación con la finalidad de seleccionar las mejores soluciones al problema a ser resuelto.

En el presente trabajo se realizó un análisis de las técnicas de agrupamiento que se apoyan del uso de algoritmos genéticos para encontrar una solución óptima al problema de agrupamiento. Esto con la finalidad de aplicar los algoritmos estudiados a los datos de calidad del aire.

Introducción

La presente tesis se centra en el área de Descubrir Conocimiento (del inglés *Knowledge Discovery*) y presenta una descripción de los algoritmos de agrupamiento que se apoyan del uso de algoritmos genéticos para encontrar el óptimo de una función objetivo.

Inicialmente se presenta el estado de arte relacionado con cambio climático y calidad del aire incluyendo conceptos que introducen al problema analizado, además se presenta el área de descubrir conocimiento y sus áreas afines.

Posteriormente se detalla el proceso de descubrir conocimiento, así como las técnicas comunes para realizar esta tarea; además se ilustra la utilidad de estas técnicas con ejemplos de aplicaciones realizadas durante el desarrollo de la tesis.

Después se detallan los algoritmos de agrupamiento analizados que usan a los algoritmos genéticos en el proceso de la creación de grupos y se detalla la evaluación de la efectividad de estos algoritmos.

Se complementa con la propuesta de un algoritmo híbrido, así como los resultados experimentales del procesamiento de los datos de calidad del aire.

Por último se detallan las conclusiones y el trabajo futuro.

Objetivos de la tesis

0.1. General

Investigar y aplicar técnicas para el descubrimiento de conocimiento y cómputo suave en las bases de datos del proyecto de cambio climático (FCC, IQ, ICUAP), con el fin de inferir conocimiento y generar modelos que sean de importancia para los especialistas en el área ambiental.

0.2. Particulares

Estudiar y analizar los algoritmos propuestos en la comunidad de cómputo suave y para descubrir conocimiento.

Implementar algoritmos para descubrir conocimiento y de cómputo suave que permitan extraer conocimiento de las bases de datos de cambio climático.

Proponer algoritmos híbridos combinando las técnicas de descubrir conocimiento y cómputo suave que extraigan patrones que permitan la toma de decisiones en el área de cambio climático.

Índice general

Índice de figuras

Índice de tablas

Capítulo 1

Estado del arte

1.1. Introducción

El ser humano a lo largo de su existencia ha ido cambiando su entorno para vivir de manera cómoda y segura, prueba de ello son los grandes alcances para transportarse por cielo, mar, tierra y el espacio. Los avances tecnológicos han facilitado los hábitos cotidianos, los negocios, la fabricación de grandes cantidades de productos, etc.
Sin embargo estos avances han tenido un efecto negativo en el medio ambiente, a tal grado que hemos terminado con muchas de las especies que compartían este planeta con nosotros, hemos avanzado tanto que al mismo tiempo estamos cavando nuestra propia tumba.
Es claro que no podemos revertir esta afectación al medio ambiente, pero si podemos disminuir en gran medida el problema que hemos ocasionado.
Si pensamos en un día normal en la vida de un ser humano, al inicio del día buscamos darnos un baño, transportarnos a la escuela o trabajo, para ver a los amigos o a la familia, comer, practicar alguna actividad de esparcimiento y finalmente dormir. Lo que no notamos durante el desarrollo de nuestras actividades, es que contaminamos nuestro planeta, si nos centramos en el aire por ejemplo al transportarnos se liberan gases que contaminan la atmósfera, además de que respiramos estos contaminantes mientras caminamos o platicamos; peor aún recibimos quemaduras en la piel por la reacción de estos gases con el sol. Es por ello que al vivir en una zona con grandes fuentes de contaminación nos vemos obligados a pagar las consecuencias.
Por lo tanto, es necesario saber cuanto contaminamos el aire y como esta

contaminación afecta a nuestra salud. Un intento para saber que tan grave es la contaminación en una ciudad es el monitoreo de la calidad del aire, esto nos permite saber que medidas debemos tomar en un instante determinado. Existen estaciones de monitoreo que a cada momento registran la actividad de los contaminantes en ciudades con gran número de población. En este momento podríamos preguntarnos ¿Qué tendencias hay en estas medidas?, ¿Existe un patrón en los registros de los contaminantes?, ¿Qué contaminante es el que se presenta con mayor frecuencia en la zona donde vivo? . Estas preguntas se encuentran sumamente involucradas con la información que se recolecta día a día, sin embargo su análisis no es tan sencillo, es por ello que necesitamos de las Ciencias de la Computación para el procesamiento de los datos.
En la presente tesis de maestría se han analizado dos áreas de las Ciencias de la Computación con el propósito de descubrir información relevante en los datos que nos permita tener mayor información para disminuir este problema; la primera área es el computo suave, en la cual se modelan principios biológicos a nivel computacional, para resolver problemas que no pueden ser solucionados por técnicas generales de la computación debido a su rigidez; la segunda área es descubrir conocimiento, que como su nombre lo indica pretende extraer información importante dentro de grandes cantidades de datos, que debido al gran número de variables e inmensidad son imposibles de analizar por el ser humano sin el uso de herramientas ad hoc.

1.2. Cambio climático

El proceso de cambio climático se perfila como el problema ambiental global más relevante de nuestro siglo, en función de sus impactos previsibles sobre los recursos hídricos, los ecosistemas, la biodiversidad, los procesos productivos, la infraestructura, la salud pública y, en general, sobre los diversos componentes que configuran el proceso de desarrollo.
El término suele usarse de forma poco apropiada, para hacer referencia tan sólo a los cambios climáticos que suceden en el presente, utilizándolo como sinónimo de calentamiento global. La Convención Marco de las Naciones Unidas sobre el Cambio Climático (CMNUCC) usa el término cambio climático sólo para referirse al cambio por causas humanas [20].

Definición 1. *Cambio climático (según la CMNUCC)*
Por "cambio climático" se entiende un cambio de clima atribuido directa o indirectamente a la actividad humana que altera la composición de la atmósfera mundial y que se suma a la variabilidad natural del clima observada durante períodos de tiempo comparables [6].

El clima es una descripción estadística de las condiciones de tiempo y sus variaciones, incluyendo condiciones promedio y extremas. El cambio climático se refiere a un cambio en estas condiciones que persiste por un periodo extendido, comunmente decadas o más.

Definición 2. *Cambio climático*
El cambio climático es un cambio en el patrón promedio del clima sobre un largo periodo de tiempo [1].

El clima tiene variables como temperatura y la variación en las precipitaciones pluviales. Estos cambios en el clima de día a día entre estaciones y de un año al siguiente, no representan cambios climáticos. El periodo para estimar un cambio es usualmente 30 años o más, que sea lo suficientemente largo para mostrar un gran cambio en el clima.
El clima puede ser definido para un lugar o región en particular, usualmente en base a los patrones de precipitación local o variaciones de temperatura estacionales. También es definido para el planeta entero, el clima global es una variable promedio de la temperatura de la superficie.
Los gases de efecto invernadero juegan un rol importante en la determinación del clima y causan el cambio climático.
Los gases de efecto invernadero incluyen vapor de agua, dióxiodo de carbono (CO_2), metano (CH_4), óxido nitroso (N_2O) y algunos gases industriales tales como clorofluorocarbonos (CFC_s). Estos gases actúan como una manta aislante, manteniedo la superficie de la tierra más caliente de lo que debería estar si estos gases no se presentaran en la atmósfera.Excepto por el vapor de agua, las concentraciones atmosféricas de estos gases son directamente generados por las actividades humanas. Una vez liberados a la atmósfera muchos de estos gases permanecen ahí por largo tiempo: en particular, una significante fracción de las emisiones de (CO_2) permanece en el sistema climático por cientos o miles de años.
Los efectos del cambio en los niveles de los gases de efecto invernadero sobre el clima pueden ser distinguidos a partir de los efectos en otros factores como

cambios en la radiacción solar. Estos factores conducen a direfentes patrones o huellas, resultado del cambio climático, los cuales asisten a identificar la causa de los cambios observados. Por ejemplo, el incremento en la radiación solar lleva a calentar la parte superior e inferior de la atmósfera y el resultado son días mas calientes que las noches. Por otro lado el incremento en los gases de efecto invernadero se refleja en un enfriamiento y no un calentamiento de la estratósfera lo que ocasiona noches mas cálidas que los días. Los patrones observados de cambio indican el incremento en los gases de efecto invernadero [1].

¿Cómo ha cambiado el clima en la tierra en un pasado distante?

El clima ha variado enormemente a través de la historia de la tierra. Desde que la tierra fue formada hace 4.5 mil millones de años, el clima ha cambiado dramaticamente, muchas veces debido a cambios en los oceanos y la separación de los continentes, variaciones naturales en los niveles de los gases de efecto invernadero en la atmosfera, la intensidad del sol y la orbita de la tierra alrededor del sol.

Evidencia del pasado muestra que el clima es sensitivo a pequeñas influencias. Durante los últimos millones de años la temperatura promedio de la superficie de la tierra ha subido y bajado en alrededor de $5°C$,a través de los 10 principales ciclos en la era de hielo. Los últimos 8000 años han sido relativamente estables hacia un aumento en el calentamiento en este rango de temperatura. Estos ciclos fueron iniciados por sutiles variaciones en la orbita de la tierra que alteraron el patrón de la absorción solar. Las medidas de los núcleos de hielo y otras fuentes sugieren fuertemente que la temperatura cambió, otros cambios fueron provocados, esto generó un efecto amplificado: durante los periodos de calor dióxiodo de carbono (CO_2) y metano (CH_4) fueron liberandos a la atmósfera, y las capas de hielo retrocedieron y regresaron menos luz solar al espacio. Esto significa que algunas pequeñas influencias fueron amplificadas a enormes cambios.

Una importante implicación de la busqueda de cambios climáticos en el pasado es que ciertos procesos similares son probablemente amplificados en la actualidad hacia el clima por influencias humanas.

Registros del pasado muestran que el clima puede cambiar abruptamente. Los cambios más graves en la temperatura global son mostrados evidentemente en la geología; esto ocurrió lentamente sobre decenas de miles o millones de años, mucho más gradualmente que el calentamiento del siglo

pasado. Sin embargo, algunos cambios rápidos han sido documentados en muchos calentamientos climáticos del pasado y más reciente eras de hielo.
Uno de estos cambios rápidos tomó lugar 56 millones de años atrás, cuando la temperatura global se incrementó por cerca de $5°C$, acompañada por una inexplicable liberación de gases de efecto invernadero a la atmosfera. Esta liberación podría haber sido tan rápida que es comparable con la acutal liberación de quema de combustibles fósiles por parte de los humanos. Otros cambios rápidos sucedieron durante la última edad de hielo, de $5°C$ o más tan solo hace algunas décadas de manera regional surgieron colapsos repentinos de glaciares o cambios en los océanos actuales.
Aunque en el milenio anterior la revolución industrial fué relativamente estable, hubo variaciones en el clima sobre este periodo. Durante el periodo cálido medieval (800-1300 d.c.) y una pequeña edad de hielo (1500-1800 d.c.) son dos bien conocidos episodios durante los pasados miles de años. El hemisferio norte estuvo $1°C$ más caliente en promedio durante el período anterior que durante el siguiente. Sin embargo, ciertas evaluaciones indican que el promedio de temperatura en el hemisferio norte en los últimos cincuenta años, ha sido mas caliente que durante el periodo del calentamiento medieval y las temperaturas sobre la última década son más calientes aún.
Los registros son escasos en el hemisferio sur, pero los poco disponibles indican escasa o ninguna relación con el calentamiento en el hemisferio norte, durante el calentamiento del periodo medieval; a diferencia de esto el enfriamiento es globalmente coherente con la pequeña edad de hielo.
Existen también variaciones regionales en el clima, particularmente precipitaciones pluviales, que no estan asociadas con los cambios globales. Por ejemplo sequías regionales parecen haber contribuido al colapso del antiguo imperio Acadio en el medio oriente y los Mayas en México.

¿Cómo ha cambiado el clima durante un pasado reciente?

El promedio global de temperaturas ha incrementado sobre el siglo pasado. Medidas de cientos de termómetros alrededor del globo terraqueo, tanto en la tierra como en el océano, muestran que el promedio cerca de la superficie se incremento sobre 100 años hasta el 2009 por más de $7°C$. Muchas de estas mediciones se iniciaron en la segunda mitad del siglo XIX, y no fueron diseñadas inicialmente para ser usadas para monitoreo ambiental. Esto quiere decir que estas tienen que ser cuidadosamente analizadas para

tratar con cambios en los instrumentos, prácticas de observación, ubicación y el crecimiento de las ciudades. Despues de contar con estos problemas, los incrementos de temperatura son mayores en los continentes interiores de Asia y el norte de África, regiones que están alejadas de las principales áreas de crecimiento de la población.

1.2.1. La contaminación del aire

La contaminación del aire o contaminación atmosférica es un problema que produce cambios climáticos en todo el mundo y afecta a la salud de millones de personas. Si bien el efecto de la contaminación del aire aún no se ha evaluado en toda su magnitud, se reconoce que el problema se presenta de distintas formas dependiendo de la situación geográfica y del nivel de desarrollo.

Se entiende por contaminación atmosférica a la presencia en la atmósfera de sustancias en una cantidad que implique molestias o riesgo para la salud de las personas y de los demás seres vivos, vienen de cualquier naturaleza, así como que pueden atacar a distintos materiales, reducir la visibilidad o producir olores desagradables. El nombre de la contaminación atmosférica se aplica por lo general a las alteraciones que tienen efectos perniciosos en los seres vivos y los elementos materiales, y no a otras alteraciones inocuas. Los principales mecanismos de contaminación atmosférica son los procesos industriales que implican combustión, tanto en industrias como en automóviles y calefacciones residenciales, que generan dióxido y monóxido de carbono, óxidos de nitrógeno y azufre, entre otros contaminantes. Igualmente, algunas industrias emiten gases nocivos en sus procesos productivos, como cloro o hidrocarburos que no han realizado combustión completa [21].

La contaminación atmosférica puede tener carácter local, cuando los efectos ligados al foco se sufren en las inmediaciones del mismo, o planetario, cuando por las características del contaminante, se ve afectado el equilibrio del planeta y zonas alejadas a las que contienen los focos emisores. Los contaminantes aéreos pueden ser clasificados de manera general en dos categorías [14]:

Contaminantes primarios son aquellos que son emitidos a la atmosfera mediante fuentes como la combustión de combustibles fósiles de plantas de energía, vehículos y producción industrial, por la combustión de biomasa para fines agrícolas o propósito de limpieza de tierras y por procesos naturales como el polvo arrastrado por el viento, actividad

volcánica y respiración biológica.

Contaminantes secundarios son formados en la atmosfera cuando los contaminantes primarios reaccionan con la luz del sol, oxigeno, agua y otros químicos presentes en el aire.

Contaminantes aéreos

Los contaminantes aéreos pueden ser encontrados en ambientes exteriores e interiores, estos pueden ser divididos en tres grupos [7]:

1. Contaminantes criterio
2. Contaminantes tóxicos en el aire
3. Contaminantes biológicos

Contaminantes criterio

Contaminantes criterio es un termino usado internacionalmente para describir contaminantes aéreos que han sido regulados y son usados como indicadores de la calidad del aire. Las regulaciones o estándares son basados en criterios relativos a la salud y/o efectos ambientales [7]. A continuación se describen cada uno de los contaminantes criterio, debido a que se trabajará en las siguientes secciones con las mediciones de estos contaminantes [8].

Ozono (O_3): No es emitido directamente en el aire, pero es creado por reacciones quimicas entre oxidos de nitrógeno (NOX) y compuestos orgánicos volatiles con la presencia de la luz del sol.
El ozono O_3 es un constituyente natural de la atmósfera, pero cuando su concentración es superior a la normal se considera como un gas contaminante.
Su concentración a nivel del mar, puede oscilar alrededor de 0.01 $mg\ kg^{-1}$.
Cuando la contaminación debida a los gases de escape de los automóviles es elevada y la radiación solar es intensa, el nivel de ozono aumenta y puede llegar hasta 0.1 kg^{-1}.
Las plantas pueden ser afectadas en su desarrollo por concentraciones pequeñas de ozono. El hombre también resulta afectado por el ozono a concentraciones entre 0.05 y 0.1 $mg\ kg^{-1}$, causándole irritación de las fosas nasales y garganta, así como resequedad de las mucosas de las vías respiratorias superiores.

Dióxido de sulfuro (SO_2): Pertenece a un grupo de gases altamente reactivos, conocido como "Oxidos de sulfuro". Las mayores emisiones de SO_2 se derivan de la combustión de combustibles fósiles en plantas de energía (73 %) y otros servicios industriales (20 %). Pequeñas fuentes de emisiones de SO_2 incluyen procesos industriales como extración de metales a partir de minerales y la quema de combustibles con alto contenido en sulfuro en locomotoras, grandes barcos, entre otros. El SO_2 se encuentra ligado a un número de efectos nocivos en el sistema respiratorio.

Dióxido de nitrógeno (NO_2): Pertenece a un grupo de gases altamente reactivos, conocido como "Oxidos de nitrogeno". NO_2 se forma rápidamente de emisiones de autos, camiones, autobuses, plantas de energía, entre otros. Además de que contribuye a la formación de ozono y partículas finas contaminantes, NO_2 esta ligado con un número de efectos nocivos en el sistema respiratorio.

Monóxido de carbono (CO): Es un gas incoloro, inoloro emitido por procesos de combustión, la mayoría de las emisiones de CO en el medio ambiente provienen de fuentes móviles. El CO puede causar efectos nocivos en la salud mediante la reducción del suministro de oxígeno a los órganos del cuerpo (como el corazón y el cerebro) y los tejidos. A niveles muy altos, el CO puede causar la muerte. Cada año, aparecen varios casos de intoxicación mortal, a causa de aparatos de combustión puestos en funcionamiento en una habitación mal ventilada.
Los motores de combustión interna de los automóviles emiten monóxido de carbono a la atmósfera por lo que en las áreas muy urbanizadas tiende a haber una concentración excesiva de este gas hasta llegar a concentraciones de 50-100 partes por millon (ppm), tasas que son peligrosas para la salud de las personas.

Partículas suspendidas (PM): También conocida como la contaminación por partículas o PM, es una mezcla compleja de partículas extremadamente pequeñas y gotitas líquidas. La contaminación por partículas se compone de ácidos (tales como los nitratos y sulfatos), productos químicos orgánicos, metales, y las partículas de suelo o polvo.
El tamaño de las partículas está directamente relacionada con su potencial de causar problemas de salud. Los gobiernos ponen especial atención por las partículas que miden 10 micrómetros de diámetro o

menos, porque esas son las partículas que pasan a través de la garganta y la nariz y entran en los pulmones. Una vez inhaladas, estas partículas pueden afectar el corazón y los pulmones y causar efectos graves para la salud. Las partículas suspendidas se dividen en dos categorías:

Partículas menores a 10 micrómetros (PM_{10}): Tales como las que se encuentran cerca de las carreteras y las industrias de polvo, son más grandes que 2.5 micrómetros y más pequeñas que 10 micrómetros de diámetro.

Particulas menores a 2.5 micrómetros (PM_{25}): Tales como las que se encuentran en el humo y la neblina, son de 2.5 micrómetros de diámetro y más pequeñas. Estas partículas pueden ser emitidas directamente por fuentes tales como los incendios forestales, o se pueden formar cuando los gases emitidos por plantas de energía, las industrias y los automóviles reaccionan en el aire.

Contaminantes tóxicos en el aire

Los contaminantes tóxicos en el aire son algunas veces referidos como "Contaminantes peligrosos en el aire". Las fuentes de estos contaminantes son los vehículos, la combustión de los combustibles sólidos, emisiones industriales y materiales como pinturas y adhesivos en edificios nuevos. Los contaminantes tóxicos tienen el potencial de causar un serio daño a la salud y al medio ambiente.

Contaminantes biológicos

Los contaminantes biológicos son otra clase de contaminantes. Ellos surgen de fuentes como la contaminación microbiológica, la piel de los animales y humanos, las plagas como las cucarachas, entre otros. Los contaminantes biológicos pueden ser transmitidos de forma aérea y pueden tener un impacto significativo en la calidad de ambientes interiores.

1.2.2. Calidad del aire

La calidad del aire es medida por la concentración de los contaminantes aéreos, es decir a mayor presencia en el aire, menor es la calidad y consecuentemente mayor es el impacto en la naturaleza y los seres humanos. [16].

Una estación de monitoreo de calidad del aire obtiene la concentración de los mayores contaminantes aéreos en un tiempo específico.

Calidad del aire en la ciudad de México

Los índices de calidad del aire (ICA) son números usados por agencias del gobierno para determinar la calidad del aire en una localidad en específico. En la ciudad de México y en la zona metropolitana del valle de México (ZMVM) la contaminación del aire es medida con el índice metropolitano de calidad del aire (IMECA). El IMECA es usado para mostrar el nivel de contaminación y el nivel de riesgo que representa a la salud humana en un tiempo determinado y así poder tomar medidas de protección. El IMECA es calculado usando las medidas de horas promedio de los químicos ozono (O_3), dióxido de sulfuro (SO_2), dióxido de nitrógeno (NO_2), monóxido de carbono (CO), partículas menores a 10 micrómetros (PM_{10}) y partículas menores a 2.5 micrómetros (PM_{25}).

Categorías

Para reportar la calidad del aire, el índice emplea cinco categorías:

Buena. Cuando el índice se encuentra entre 0 y 50 puntos IMECA, la calidad del aire se considera como satisfactoria y la contaminación del aire tiene poco o nulo riesgo para la salud.

Regular. Cuando el índice se encuentra entre 51 y 100 puntos IMECA, la calidad del aire es aceptable, sin embargo algunos contaminantes pueden tener un efecto moderado en la salud para un pequeño grupo de personas que presentan una gran sensibilidad a algunos contaminantes.

Mala. Cuando el índice se encuentra entre 101 y 150 puntos IMECA, algunos grupos sensibles pueden experimentar efectos en la salud. Hay algunas personas que pueden presentar efectos a concentraciones menores que el resto de la población, como es el caso de personas con problemas respiratorios o cardíacos, los niños y ancianos. El público en general puede no presentar riesgos cuando el IMECA está en este intervalo.

Muy mala. Cuando el índice se encuentra entre 151 y 200 puntos IMECA, toda la población experimenta efectos negativos en la salud. Los miembros de grupos sensibles pueden presentar molestias graves. En este

intervalo se activan las Fases de Precontingencia y Contingencia Fase I del Programa de Contingencias Ambientales Atmosféricas (PCAA) del Valle de México.

Extremadamente mala. Cuando el valor del índice es mayor a 201 puntos IMECA, la población en general experimenta molestias graves en la salud.

1.3. Descubrir conocimiento

En las Ciencias de la Computación existe un área conocida como descubrir conocimiento (del inglés *Knowledge Discovery*); como su nombre lo dice, su intención es la de usar técnicas, que nos permitan extraer información relevante de conjuntos de datos. Es por ello que consideramos que el análisis de técnicas para descubrir conocimiento, nos va a dar el soporte necesario para analizar los datos de calidad del aire y observar que información relevante podemos encontrar en estos.

Definición 3. *Descubrir conocimiento (Knowledge Discovery)*
Descubrir conocimiento es un proceso semi automático para extraer información útil de colecciones de datos que son demasiado grandes como para ser investigados manualmente [12].

La información regresada por el proceso de descubrimiento usualmente toma la forma de patrones recurrentes o explicativos que son usualmente referidos como modelos; hay muchos tipos de modelos, por ejemplo, tenemos modelos que son representados como reglas if-then-else, así como modelos que implementan redes neuronales artificiales. Todos los modelos tienen la propiedad deseable de que tienden a ignorar detalles innecesarios y resumen las principales tendencias en los datos. Un modelo puede representar o resumir terabytes de datos y por lo tanto, facilita el acceso a la información o el conocimiento oculto en grandes cantidades de datos.
Un termino usualmente asociado con descubrir conocimiento es *minería de datos (data mining)*. La minería de datos puede ser considerada como una forma de descubrir conocimiento, esta tiene como objetivo extraer información de bases de datos; la minería de datos es usualmente referida como descubrir conocimiento en base de datos *(knowledge discovery in databases (KDD))* [12].

El descubrimiento del conocimiento es un área altamente interdisciplinaria, ya que cubre un gran rango de actividades como son el dominio de análisis, la limpieza de datos, y visualización para la evaluación y desarrollo de modelos (Figura 1.1). Sin embargo el núcleo del proceso de descubrir conocimiento es la creación de algoritmos que realicen algún tipo de reconocimiento de patrones y construyan modelos a partir de los datos encontrados.

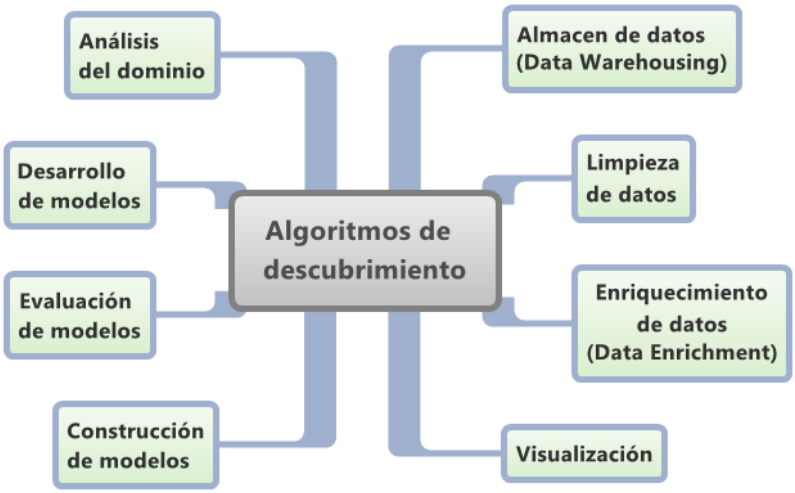


Figura 1.1: Procesos para descubrir el conocimiento.

1.3.1. Entornos para descubrir conocimiento

Un entorno o herramienta de trabajo para descubrir conocimiento debe soportar aspectos computacionales del proceso de descubrimiento.

Aspectos computacionales para descubrir conocimiento

Mencionamos en la definición 3 que el descubrir conocimiento es un proceso semiautomatizado; esto significa que es un proceso que recae fuertemente en el uso de herramientas computacionales, pero la guía de un analísta es indispensable. El analísta prueba un modelo experto y formula la tarea de descrubrimiento de tal manera que pueda ser abordado usando herramientas computacionales. Por lo tanto el analísta toma decisiones acerca de cuando un modelo es apropiado y cuando falla al resumir los datos en algúna consulta

útil o inútil. Los aspectos que requieren la intervención del analísta, especialmente el dominio de análisis, son dificiles de formalizar y automatizar, haciendo la cooperación entre el analísta y la computadora absolutamente necesaria para crear proyectos para descubrir conocimiento de manera exitosa. Las herramientas computacionales hacen posible el análisis de grandes cantidades de datos, acontinuación se presenta una descripción de las características de este tipo de herramientas.

Acceso a datos

Son herramientas para descubrir conocimiento que deben proveer una manera eficiente de accesar a los datos, por ejemplo: poder importar una tabla de datos o tener la opción de realizar consultas SQL directamente a una base de datos o a un almacén de datos. Algunas herramientas que asisten para descubrir conocimiento se encuentran embebidas en un motor de base de datos para minimizar los problemas del acceso a los datos.

Visualización

La visualización de datos es una manera poderosa de obtener conocimiento de los datos. Muchos analístas usan la visualización de datos para "tener la corazonada" de los datos e identificar la calidad de estos. Por ejemplo, un analísta podría desear observar si los datos tienen valores ausentes o si tal vez alguno de los atributos se encuentran sesgados. Otro cuestionamiento que es importante en la construcción de un modelo, es si alguno de los atributos independientes está altamente correlacionado, ya que en algunos casos, los atributos independientes altamente correlacionados pueden reducir la efectividad de los algoritmos de descubrimiento. Muchas de estas preguntas son fácilmente contestadas usando la visualización.

Manipulación de datos

Por supuesto que no es suficiente con leer, escribir y visualizar los datos; también se necesitan herramientas para manipular los datos. El enfoque de la manipulación de datos cae en una de dos categorías. En el *enfoque orientado a los atributos* podemos manipular columnas completas de una tabla de datos. Esto es particularmente útil cuando nuestro objetivo es enriquecer la tabla con información adicional agregando o eliminando columnas que representan a los atributos que consideramos inútiles para el proceso de descubrimiento.

En el enfoque *orientado a la observación* nos enfocamos en las filas de las tablas de datos. Esto es útil para remover observaciones que son defectuosas y son consideradas como valores atípicos.

Construcción de modelos y evaluación

En el corazón del proceso de la búsqueda de conocimiento usualmente encontramos dos clases de algoritmos de descubrimiento: algoritmos de aprendizaje artificial y técnicas estadísticas. Los algoritmos de aprendizaje artificial fueron desarrollados en el área de la inteligencia artificial que se remonta a finales de los años 50 y fueron diseñados para dotar de inteligencia a agentes autónomos. Las técnicas estadísticas fueron desarrolladas en el contexto de la probabilidad y la medida de la teoría al final del siglo XIX. Sin embargo, fue a finales de los años 80 y principios de los años 90 que los investigadores reconocieron que ambas areas estaba tratando con problemas similares. Con la llegada de la computación estadística, los bordes entre estas disciplinas desaparecieron y las técnias estadísticas que se ocupan de la construcción de modelos y la inferencia son casi indistinguibles del aprendizaje maquina y viceversa. Pero existe aún una diferencia entre los dos enfoques que tiene que ver principalmente con la suposición de los conjuntos que admiten durante el análisis y la construcción de modelos. Muchas técnicas estadísticas confian en el hecho de que hay una distribución normal en cualquiera de los datos o un error de modelado. Por su parte los algoritmos de aprendizaje artificial, en general no hacen estas suposiciones y por lo tanto son capaces de proveer modelos más precisos en situaciones donde las supocisiones de normalización no son garantizadas. Por otro lado, las nuevas técnicas de computación estadística como bootstrap también predicen muchas suposiciones de normalidad, una vez más desvaneciendo la diferencia entre aprendizaje máquina y estadística.
Dadas las pequeñas diferencias entre aprendizaje artificial y técnicas estadísticas, es fácil para el usuario tomar el algoritmo que responda mejor para un problema en particular. Por otro lado, algunas veces los enfoques son impuestos a los usuarios, debido a restricciones externas. Por ejemplo para una actividad de descubrir conocimiento a mano, podría ser de gran importancia que los modelos sean transparentes, es decir, que los modelos puedan facilmente ser leídos y entendídos por un ser humano, forzando al analísta a usar algo parecido a árboles de decisión o listas de reglas como modelos. Por el contrario un análisis detallado del error de modelado y técni-

cas de estadística complejas podrían ser importantes en torno a favorecer enfoques estadísticos más precisos.

Desarrollo de modelos

El desarrollo de modelos es altamente dependiente del dominio. En algunos casos esto significa simplemente predecir el valor del atributo objetivo para un conjunto de objetos. En otros casos esto quiere decir construir una aplicación entera alrededor del modelo. Considere una aplicación de puntuación de credito para un banco de hipotecas el cual tiene un modelo incluido. En un escenario típico un empleado del banco ingresa la información personal del cliente,como edad, ingresos y otras cantidades pendientes de préstamos, entonces presiona un botón. En este punto es donde la aplicación usa el modelo embebido para predecir si el cliente califica o no para un crédito hipotecario [12].

1.4. Reconocimiento de patrones

En muchas investigaciones es importante saber si existe alguna tendencia dentro de los datos o comportamiento que nos brindara información adicional que no es tan evidente por la gran cantidad de datos. El área de reconocimiento de patrones nos provee de técnicas computacionales que han demostrado ser exitosas y que permiten hacer un análisis de las características de los datos con los que se esta trabajando y de este modo poder clasificar la información que se tiene almacenada.
Según algunos autores, el objetivo básico de todas las ciencias es el reconocimiento de patrones [3].
El reconocimiento de patrones es un área de la ciencia muy general debido a que su objetivo, es encontrar estructuras en conjuntos de datos. Se puede definir de forma sencilla al reconocimiento de patrones como la búsqueda de estructuras en los datos [4]. Esta definición tiene dos implicaciones directas:

- Es un proceso necesario en muchas líneas de investigación científica.
- Es, por su propia naturaleza, una ciencia inexacta, ya que puede admitir muchas aproximaciones, bien complementarias, bien contradictorias, para llegar a una solución a un problema dado.

Entre las áreas de aplicación del reconocimiento de patrones se encuentran:

Interacción Humano Computadora: detección de voz automática, tratamiento de imágenes, procesamiento de lenguaje natural.

Defensa: reconocimiento automático de objetivos, guía y control de armamento.

Medicina: diagnósticos, análisis de imágenes y clasificación de enfermedades.

Diseño de vehículos: automóviles, aeroplanos, trenes y barcos.

Aplicaciones policiales: detección de escritura, huellas digitales y análisis de fotografías.

Estudio de recursos naturales: agricultura, geología y recursos forestales.

Industria: diseño asistido por computadora, pruebas y control de calidad.

El reconocimiento de patrones consta de varias actividades. Estas son:

- Elegir el formato de la información, buscando unas características que representen cada dato del proceso.
- Analizar las características, de forma que se puedan eliminar las no significativas.
- Agrupar los datos caracterízados, etiquetando los subgrupos naturales y homogéneos que se encuentren en el espacio de características.
- Por último, diseñar un clasificador, capaz de etiquetar cualquier punto del espacio de características.

La información necesaria para realizar sistemas de reconocimiento de patrones puede ser [15]:

- Numérica, donde se hablaría de un Sistema de Reconocimiento de Patrones Numéricos (SRPN).
- Estructural o Sintáctica.

1.4.1. Fases del reconocimiento de patrones

Hay cuatro fases en las que se puede dividir el sistema de reconocimiento de patrones. Dichas fases son:

1. *Descripción del proceso*
   En este paso se debe elegir como se va a procesar la información. Aquí es donde se elige el formato de la información (por ejemplo, un formato numérico, sintáctico o basado en reglas). Lo más habitual es el SPRN, es decir, se utiliza una lista ordenada de características, denominada vector, para representar a los datos. De esta forma, los datos estarían representados por un conjunto X tal como:

   $$X = \{x_1, x_2, ..., x_n\}$$

   Por lo tanto, X es un conjunto de n vectores de características en el espacio de características R^p (donde p es el número de características de cada objeto). Cada objeto i tendrá su vector x_i donde cada x_{ij} es el valor numérico de la característica j del objeto i.
   Por último, una distinción importante es que los datos estén etiquetados o no etiquetados. Los datos están etiquetados si se conoce la clase a la que pertenece cada vector de datos,mientras que estarán no etiquetados si no se conocen.

2. *Análisis de Características* En este paso se explora y mejora los datos recogidos en la primera fase. Los métodos que se suelen incluir son el escalado de los datos, su normalización, la representación visual de dichos datos para eliminar características redundantes o no significativas, etc. El objetivo principal de este paso es el de comprimir el espacio de características a $R^{p'}$, donde $p' < p$.

3. *Análisis de Agrupaciones*
   A esta fase se llega con un conjunto de datos, descritos en la primera fase y comprimidos en la segunda. El objetivo es el de asignar etiquetas a los objetos que identifiquen a los subgrupos naturales y homogéneos del conjunto total de objetos. Este problema se denomina agrupamiento y sus características principales son:

   - Los datos suelen estar no etiquetados.
   - No se conocen las etiquetas de los subgrupos buscados.

- Además, el número de subgrupos puede ser desconocido.

En el caso de los SRPN, hay varios tipos de algoritmos que pueden resolver el problema. Una primera clasificación de los algoritmos de agrupamiento podría ser [15]:

Por el tipo de modelo de algoritmo: determinístico, probabilístico o borroso.

Por el dominio del algoritmo: global o local.

Por el tipo de criterio del algoritmo: jerárquico, función objetivo, en forma de grafos.

Por el tipo de algoritmo: iterativo, aglomerativo o de descomposición.

Por la arquitectura: en serie, en paralelo o híbrida.

4. *Diseño del Clasificador*
   Otro problema, potencialmente más ambicioso que el agrupamiento, es el de la clasificación. Se denomina así al hecho de partir el propio espacio de caracteristicas R^P. La diferencia entre agrupamiento y clasificación es que, en el primer caso, el agrupamiento sólo etiqueta a un conjunto de datos $X \subset R^P$ mientras que el clasificador puede etiquetar cualquier punto en el espacio entero de características R^P.
   Es común, pero no necesario, que los clasificadores se diseñen con datos etiquetados [1]. Las funciones que se utilizan para realizar la partición del espacio van desde funciones implícitas, tales como perceptrones multicapa o reglas del vecino más cercano a funciones explícitas, tales como funciones discriminantes o reglas del prototipo más cercano.

1.5. Cómputo suave

Los seres humanos tenemos la habilidad de razonar y tomar decisiones diariamente para realizar tareas fundamentales que nos permiten interactuar con nuestro ambiente y con otras personas. Este tipo de habilidad no es compartida, en muchos casos, por sistemas automáticos.
La pericia que los humanos emplean para, por ejemplo, conducir un automvil

[1] Lo que se denomina "Aprendizaje supervisado"

en forma segura, desarrollar planes para lograr ciertos objetivos, coordinar nuestras actividades con otros seres humanos, o comprender el contenido de una novela no son, en este momento, emuladas eficientemente por sistemas automáticos [17].
Actualmente un tema estudiado por muchos investigadores, es el cómputo suave o Soft Computing. Su objetivo es bien concreto: aumentar el "coeficiente intelectual" de las computadoras dándoles la habilidad de imitar a la mente humana, la cual es blanda, suave, flexible, adaptable e inteligente. En palabras de Lotfi Zadeh, Profesor de la Universidad de California y reconocido experto mundial en la materia, "es la antítesis de la computación actual, asociada con la rigidez, la fragilidad, la inflexibilidad y la estupidez".

Definición 4. *Cómputo suave (Soft Computing)*
Cómputo suave o flexible (del inglés Soft Computing) es el nombre por el que se conoce a un conjunto de metodologías (basadas en ideas inspiradas por la biología, psicología, y lingüística) que buscan la solución a tales problemas, caracterizados por la necesidad de interactuar eficientemente con sistemas complejos cuando la información disponible es insuficiente.

Esta área de la computación se formaliza a inicios de los 90's [24], las técnicas que conforman esta área son [23]:

- Redes neuronales
- Sistemas difusos
- Cómputo bioinspirado (Computación evolutiva, Metaheurísticas)
- Probabilidad (Redes bayesianas)
- Teoría del caos

Las técnicas de cómputo suave se asemejan más a los procesos biológicos que a las técnicas matemáticas tradicionales, que se basan en sistemas formales. Además las técnicas de cómputo suave intentan complementarse unas a otras, explotan la tolerancia de la imprecisión, la verdad parcial y la incertidumbre para un problema específico. Como lo señala Lofti A. Zadeh (1994) [24]:

Definición 5. *Cómputo suave*
Cómputo suave no es un cuerpo homogéneo de conceptos y técnicas, más bien es una mezcla de distintos métodos que de una forma u otra cooperan desde sus fundamentos.

Los métodos de la computación dura no proveen de suficientes capacidades para desarrollar e implementar sistemas inteligentes. En lugar de confiar en las habilidades del programador, un verdadero programa de computación suave aprenderá de su experiencia por generalización y abstracción, emulando la mente humana tanto como pueda, especialmente su habilidad para razonar y aprender en un ambiente de incerteza, imprecisión, incompletitud y verdad parcial, propios del mundo real. De esta forma, es capaz de modelizar y controlar una amplia variedad de sistemas complejos, constituyéndose como una herramienta efectiva y tolerante a fallas para tratar con los problemas de toma de decisiones en ambientes complejos, el razonamiento aproximado, la clasificación y compresión de señales y el reconocimiento de patrones. Sus aplicaciones están relacionadas, entre otras, con el comercio, las finanzas, la medicina, la robótica y la automatización.

1.6. Aportación

En la presente tesis se aplican metodologías para descubrir conocimiento, se ha partido de la idea de usar como base técnicas de agrupamiento con la finalidad de analizar las mediciones de los contaminantes criterio de calidad del aire del Distrito Federal y la Zona Metropolitana del Valle de México (ZMVM). Las técnicas de agrupamiento forman parte de la clasificación no supervisada en el área de aprendizaje artificial, específicamente se han estudiando técnicas de agrupamiento que emplean algoritmos genéticos. Los algoritmos genéticos forman parte del cómputo suave, la idea de usar algoritmos genéticos se debe a que ellos pueden acelerar la obtención de resultados. Es importante tener una comparación del desempeño de los algoritmos genéticos en el agrupamiento contra las técnicas clásicas (por ejemplo, algoritmos jerárquicos), es así como se ha tenido que evaluar la eficiencia y efectividad de estos algoritmos, para identificar si verdaderamente estos proveen una mejora y alcanzan la solución óptima a un problema dado.
La aportación principal radica en el empleo de los algoritmos de agrupamiento para el descubrimiento de patrones en los datos (conocimiento), de este modo ayudar a los tomadores de decisiones a tener información que les puedan servir para comprobar si las medidas que están aplicando son las correctas o necesitan ser modificadas.
En las siguientes secciones se describen con mayor detalle las técnicas estudiadas, una aclaración importante es que estas técnicas pueden ser encontradas

con diferentes nombres en español debido al que son áreas y técnicas con pocos años en las Ciencias de la computación y la variedad en los nombres se debe a que son traducciones del inglés; se ha decidido usar el nombre descubrir conocimiento para knowledge discovery, aprendizaje artificial para machine learning y cómputo suave para soft computing.
El capítulo 2 describe las técnicas para descubrir conocimiento estudiadas; en el capítulo 3 se describen los algoritmos genéticos, técnica del computo suave sobre la que se basan los algoritmos de agrupamiento estudiados, así como la evaluación de la efectividad de dichos métodos y las conclusiones de estos; el capítulo 4 presenta la propuesta de un algoritmo híbrido con la finalidad de analizar si este mejora los resultados de los algoritmos estudiados, ademas se presentan los resultados experimentales al problema de calidad del aire. Finalmente en el capítulo 5 se presentan las conclusiones y el trabajo a futuro.

Capítulo 2

Técnicas para descubrir conocimiento

2.1. Aprendizaje artificial

El objetivo del aprendizaje artificial (en inglés *machine learning*) es generar un modelo adecuado para un proceso de etiquetado que se aproxima al proceso original lo más fielmente posible. El término aprendizaje artificial también suele ser referido como aprendizaje automático.

Definición 6. *Aprendizaje artificial [12]*
Sean:

- *Un universo de datos* X
- *Un conjunto muestra S, donde* $S \subset X$
- *Alguna función objetivo (proceso de etiquetado)* $f : X \to \{true, false\}$
- *Un conjunto de entrenamiento de etiquetado D, donde* $D = \{(x, y) \mid x \epsilon S \; y \;\; y = f(x)\}$

Construir una función $\hat{f} : X \to \{true, false\}$ *usando D tal que* $\hat{f}(x) \cong f(x)$ *para todos los* $x \epsilon X$

El universo de datos X es el conjunto de objetos de interés; el conjunto de muestra S es un subconjunto del universo de datos, el cual actúa como una población representativa del universo de datos con el fin de hacer que el

proceso de construcción de modelos sea posible. La función objetivo f es el proceso que proporciona las etiquetas observables. Se supone que f es capaz de proporcionar un valor adecuado en verdadero, falso para algun elemento de X, cuando el elemento es observado.
Utilizamos esta propiedad de la función objetivo de construir el conjunto de entrenamiento D mediante la observación de las etiquetas de los objetos en el conjunto de muestras S. El aprendizaje artificial hace uso de datos de entrenamiento etiquetados, esto se conoce como aprendizaje supervisado. Hay otro tipo de aprendizaje artificial referido como aprendizaje no supervisado, en este tipo no hay necesidad de etiquetar los datos de entrenamiento. Finalmente en nuestra definición el aprendizaje artificial puede ser visto como calcular la función $\hat{f}$ como una aproximación a un modelo del proceso f basado en los ejemplos de entrenamiento en D. Es decir, el resultado del aprendizaje artificial es un modelo de la función de etiquetado original. Sin embargo, por conveniencia a menudo decimos que $\hat{f}$ es un modelo de los datos de entrenamiento D.
Los nombres de las etiquetas true,false son arbitrarios, en lugar de true y false podríamos haber utilizado T y F, 0 y 1, o azul y verde. El hecho importante es que este conjunto contiene dos etiquetas distintas: una para la clase a la que pertenece y otra para la clase que no pertenece. También podemos considerar los problemas de clasificación con más de dos posibilidades. La única diferencia con nuestra definición de aprendizaje artificial anterior sería que el codominio de la función f y su modelo $\hat{f}$ es un conjunto que incluye un número apropiado de etiquetas distintas.
Una vez que tenemos un modelo del proceso de etiquetado original, dos cosas interesantes que puede lograrse. En primer lugar, podemos utilizar el modelo para calcular o predecir la etiqueta de un elemento en el universo de datos X sin tener que observar este elemento. En segundo lugar, el modelo puede proporcionar alguna información sobre el proceso de etiquetado original. Es decir, un modelo posee cierta capacidad explicativa.

2.1.1. Estructura de los datos

Los objetos a clasificar pueden ser tan variados como el universo de datos, usualmente son descritos por una colección de características o atributos.
Un data set es una colección de datos, usualmente presentada en forma tabular. Cada columna representa una variable en particular; cada fila corresponde a un miembro (conocido también como objeto) del conjunto de datos en

cuestión; un data set tiene varias características las cuales definen su estructura y propiedades, estas incluyen el número y tipo de atributos o variables así como medidas estadísticas como la desviación estándar. Los valores pueden ser números reales o enteros, para representar la altura de una persona en centímetros por ejemplo, o también datos nominales (es decir que no son valores numéricos), por ejemplo el nivel de estudios de una persona.
Existe un repositorio de data sets para machine learning en:

archive.ics.uci.edu/ml/index.html

donde se pueden encontrar conjuntos de prueba para analizar el comportamiento de algún algoritmo. Un ejemplo es el iris data set, el cual fue introducido por Ronald Fisher en 1936 [10].

Iris data set

Este data set consiste de cincuenta muestras de cada una de tres especies de la flor de iris (iris setosa, iris virginica e iris versicolor), las cuales se pueden observar en la figura 2.1; se midieron cuatro características para cada muestra, las cuales son la longitud y ancho del sépalo y pétalo en centímetros. Basado en la combinación de las cuatro características, Fisher desarrolló un modelo discriminante lineal para distinguir cada una de las especies [22]. Basandose en el modelo del discriminante lineal de Fisher, este data set se

(a) Iris virginica. (b) Iris versicolor. (c) Iris setosa.

Figura 2.1: Especies del iris data set.

ha convertido en un caso de pruebas para muchas técnicas de clasificación en machine learning. Sin embargo el uso del data set en clustering no es común ya que el data set contiene dos clusters con una separación obvia; uno de los clusters contiene a las especies de iris setosa, y el otro contiene las especies de

iris virginica e iris versicolor las cuales no son separables sin la información de Fisher. Esto convierte al data set en un buen ejemplo para explicar la diferencia entre la clasificación supervisada y no supervisada [22].
En la tabla 2.1 se muestra un subconjunto de los datos de la flor de iris, por su parte en la tabla 2.2 se presenta un resumen estadístico del data set.

Longitud de sépalo	Ancho de sépalo	Longitud de pétalo	Ancho de pétalo	Especies
5.1	3.5	1.4	0.2	Setosa
6.5	2.8	4.6	1.5	Versicolor
4.7	3.2	1.3	0.2	Setosa
7.7	3.8	6.7	2.2	Virginica

Tabla 2.1: Subconjunto del universo de datos del iris data set.

Atributo	Min	Max	Media	Desviación estandar
1	4.3	7.9	5.84	0.83
2	2.0	4.4	3.05	0.43
3	1.0	6.9	3.76	1.76
4	0.1	2.5	1.20	0.76

Tabla 2.2: Resumen de los atributos del iris data set.

2.1.2. Representaciones del modelo

Desde que nosotros deseamos la aproximación de la función $\hat{f}$ a la función objetivo f sea modelada, estamos interesados en la representación apropiada de los modelos $\hat{f}$. Típicamente se consideran dos tipos de representaciones para los modelos [12]:

1. Representación transparente (o modelos tranparentes).
   - *a*) Reglas if-then-else.
   - *b*) Árboles de decisición.
2. Representación no transparente (o modelos no tranparentes).
   - *a*) Los pesos en las conexiones entre los elementos de una red neuronal.

b) La combinación lineal de vectores en las maquinas de soporte vectorial.

Los modelos transparentes son representaciones que pueden ser interpretadas por los humanos sin ayuda; por su parte los modelos no transparentes no pueden ser interpretados sin ayuda.
La representación de modelos es un tema importante porque este dicta que tan bien se pueden modelar ciertas funciones objetivo.
Si consideramos una tabla de datos como una representación de un modelo; la representación del modelo revisa el conjunto de entrenamiento y memoriza todos los objetos. Esto significa que el modelo tendrá perfecto conocimiento de los objetos en el conjunto de entrenamiento pero fallará al producir resultados significativos con los objetos que no estén en la tabla.
Los algoritmos que dan lugar a sofisticadas respresentaciones de un modelo descubren regularidades en los objetos relativos a sus correspondientes etiquetas, estas regularidades son entonces codificadas en apropiadas representaciones del modelo.
Es interesante observar que, en general las representaciones tranparentes de los modelos fallan en desempeño comparadas con las representaciones no transparentes del modelo. La restricción de que un modelo sea interpretable por la gente sin necesidad de ayuda parece interferir con los procesos de modelado, en donde un modelo transparente no es capaz de clasificar ciertos fenómenos tan efectivamente como lo hacen los modelos que no son transparentes.

2.2. Técnicas del aprendizaje artificial

2.2.1. Árboles de decisión

Introducción

Los arboles de decisión son uno de los paradigmas más utilizados en el mundo del aprendizaje artificial. La sencillez del modelo, la accesibilidad a diferentes implementaciones, la explicación que aporta a la clasificación, la posibilidad de ser representados gráficamente, la rapidez a la hora de clasificar nuevos patrones, etc. Son factores que han influido en su difusión.
Los árboles de decisión entran dentro de los métodos de clasificación supervisada, es decir, tendremos una variable dependiente o clase, y el objetivo

del clasificador va a ser averiguar dicha clase para casos nuevos. La construcción del árbol de clasificación se realiza mediante un proceso de inducción, de ahí que también sean denominados como Top-Down-Induction-Decision-Trees (TDIDT) [18].

Descripción general de los árboles de decisión

Un clasificador puede ser definido como una función $d(x)$ definida en el espacio de clasificación X, que relaciona a cada patrón o ejemplo x del espacio de clasificación con una clase del conjunto de posibles valores a los que puede pertenecer $C_m\,(m = 1, \cdots, M)$. Otra posible definición, más cercana al paradigma de árboles de decisión, sería la siguiente: un clasificador es una partición del espacio de clasificación X en M subconjuntos disjuntos $A_1, A_2, \cdots, A_M$, siendo X la unión de todos ellos y para todo x perteneciente a A_m la clase predicha es C_m. Esta segunda definición se aproxima más al mecanismo de funcionamiento de los árboles de decisión ya que lo que hacen es dividir el espacio de clasificación en zonas, de manera que a los patrones que pertenecen a cada zona se les asigna una de las posibles clases.
En la figura 2.2 se muestra un árbol de decisión simple. Todo árbol de decisión comienza con un nodo al que pertenecen todos los casos de la muestra que se quiere clasificar. Se le denomina nodo raíz y en la figura aparece en negro. El resto de los nodos se dividen en nodos intermedios no terminales (en la figura aparecen en color blanco), y nodos hoja o terminales, es decir, nodos que no se van a dividir más (en la figura aparecen de color gris). En la fase de construcción del árbol cada nodo hoja se hace corresponder con una categoría concreta de la variable clase. De esta manera, los nodos hoja representan las diferentes particiones en las que se ha dividido el espacio de clasificación. Los nodos que 'cuelgan' de un nodo concreto se dice que son nodos hijo de dicho nodo, y al nodo del que parten las flechas o ramas se le denomina nodo padre. A la hora de clasificar cada patrón, el punto de partida es el nodo raíz y, dependiendo de los valores de la variable predictora por la que se pregunta, los casos se van distribuyendo por los nodos hijo (se dice que 'caen' en dichos nodos). El proceso se repite en cada nodo hasta llegar a los nodos hoja. En este ejemplo, a todos los casos de la muestra a clasificar se les pregunta por el valor que tienen en la variable continua X_2. Los casos que tengan un valor menor o igual que w irán al nodo intermedio izquierdo, y los que tengan un valor mayor que w irán al nodo de la derecha. Estos últimos casos quedarán clasificados con la clase que en la fase de entrenamiento se le

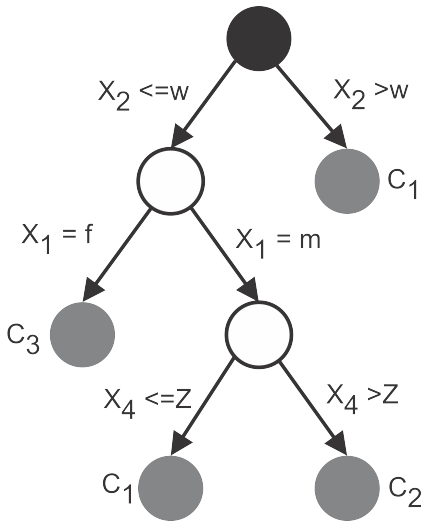


Figura 2.2: Ejemplo de un árbol de decisión.

ha asociado a dicho nodo hoja (C_1), y se dice que han 'caído' en dicho nodo hoja. A los casos que les ha correspondido el nodo intermedio se les vuelve a hacer otra pregunta, en este caso sobre la variable discreta X_7 (que podría ser, por ejemplo, el sexo de una persona: masculino o femenino). Como consecuencia de esa pregunta , algunos casos quedarán clasificados como C_3, y al resto de los casos se le volverá a realizar otra pregunta, en este caso sobre la variable continua X_4, para quedar finalmente clasificados como C_1 o como C_2.

Como se puede ver, pueden existir diferentes nodos hoja etiquetados con la misma clase. Al número de nodos hoja que tiene un árbol se le suele denominar complejidad del árbol. Para el caso del ejemplo, la complejidad tendría un valor de 4 ya que se ha dividido en cuatro partes el espacio de clasificación.

Como ya se ha comentado anteriormente, uno de los motivos por los que se utiliza ampliamente este paradigma es por el hecho de aportar una explicación a la clasificación. Si nos fijamos en el ejemplo de la figura 2.2, todos los casos que han sido clasificados como C_3 tienen en común que su valor en X_2 es menor que w y que el valor de X_7 es f. Esto, que es aplicable al resto de los nodos hoja, proporciona el motivo de porque se han clasificado como C_3. Esta capacidad explicativa es algo que, por ejemplo, una red neuronal artificial no es capaz de proporcionar. Resumiendo, el conjunto de preguntas que se encuentran en el camino que va desde el nodo raíz hasta

el nodo hoja en el que ha caído el patrón proporciona una explicación de la clasificación realizada. Este aspecto es muy importante en muchos ámbitos: diagnostico en medicina, detección de fraude, campañas de marketing, etc. Ya que proporciona información añadida que puede ser utilizada por el médico para conocer los síntomas, el investigador de la compañía de seguros para conocer el perfil del cliente supuestamente fraudulento, el departamento de marketing de una empresa que quiere mantener ciertos clientes, etc. Se puede decir que los árboles de decisión, además de clasificar, son capaces de extraer una estructura que representa, en cierta medida, el concepto o el patrón de comportamiento que hay asociado a la muestra sobre la que se ha inducido.
El proceso de construcción de un árbol de decisión comienza por el nodo raíz, el cuál tiene asociados todos los patrones o casos de entrenamiento. Lo primero que hay que hacer es un análisis para determinar cual es la variable por la que hay que preguntar para dividir la muestra de entrenamiento original (nodo raíz) en un conjunto de particiones, nodos hijo, buscando que en los subconjuntos generados haya una mínima variabilidad respecto a la clase. El caso ideal sería encontrar la variable sobre la que haciendo la pregunta adecuada, poder distribuir todos los casos de manera que en cada nodo hijo solo existieran casos pertenecientes a una sola clase. Evidentemente, hacer esto en una única pregunta es muy poco probable, de ahí que el proceso de construcción sea recursivo, es decir, que una vez que se haya determinado cual es la variable con la que se obtiene la mayor homogeneidad respecto a la clase en los nodos hijo, se vuelve a realizar el análisis para cada uno de los hijos buscando nuevas particiones que consigan una mayor homogeneidad. Aunque en el límite, el proceso se pararía cuando todos los nodos hoja contuvieran casos de una única clase, no siempre es deseable llegar al extremo de que, para decidir que un nodo sea hoja, es decir, dejar de desarrollar el árbol en ese nodo tengan que ser absolutamente todos los casos de la misma clase.
Hay que tener en cuenta que como paradigma de clasificación supervisada que es, tendremos un conjunto de entrenamiento para construir el árbol, y, aunque dicho conjunto debe de ser una muestra representativa de la distribución real, muchas veces esto no es del todo posible, y nos interesa que el árbol se ajuste (se fije) algo menos en la muestra de entrenamiento para que luego no cometa demasiados errores sobre la muestra de prueba (problema de sobre entrenamiento o sobre ajuste). Este aspecto es muy importante en el caso de los árboles y veremos diferentes estrategias a la hora de afrontarlo.
El algoritmo 1 presenta de manera general los pasos para procesar árboles

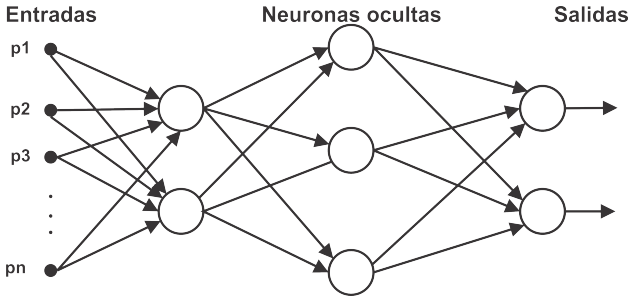
de decisión.

Algoritmo 1 Algoritmo general de inducción de árboles de clasificación

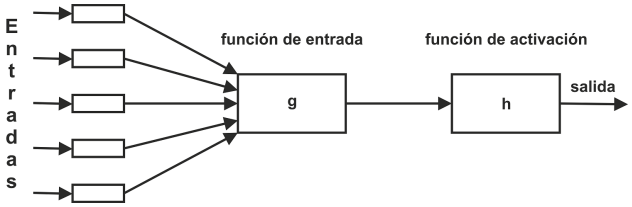
Entrada: **Conjunto de casos de entrenamiento, E, con sus variables y su clase**
Salida: **Árbol de decisión (T)**
Inicio
  si todos los casos en E son de la misma clase C_j **entonces**
    Resultado nodo simple etiquetado como C_j
  si no
    Seleccionar un atributo X_i con valores $X_{i1}, \cdots, X_{il}$
    Particionar E en $E_1, \cdots, E_l$ de acuerdo a los valores de X_i
    Construir subárboles $T_1, \cdots, T_l$ para $E_1, \cdots, E_l$
    El resultado final es un árbol con raíz X_i y subárboles $T_1, \cdots, T_l$
    Las ramas entre X_i y los subárboles están etiquetados mediante $x_{i1}, \cdots, x_{il}$
  Fin
Fin

2.2.2. Redes neuronales artificiales

El esquema general de una red neuronal típica puede apreciarse en la figura 2.3 inciso a. La red es un operador no lineal formado por un conjunto de elementos simples de procesamiento o neuronas, que se conectan entre sí y con el medio externo a través de conexiones o sinapsis, que poseen un peso asociado. La figura 2.3 inciso b, ilustra el modelo habitualmente utilizado para una neurona artificial. Esencialmente, la operación de la neurona involucra el cálculo de una función de entrada h, a partir de las señales que ingresan a la misma, y la posterior aplicación de una función de activación g, en general no lineal. En algunos casos, existe una entrada umbral (offset), que puede ser siempre asociada a una entrada adicional de valor 1 y peso igual al umbral. La salida de cada neurona dependerá, entonces de sus señales de entrada, el peso asociado a cada entrada y de las características de las funciones de entrada y activación. En la mayoría de los casos, los elementos plásticos de la red son los pesos de las conexiones, y el mecanismo utilizado para adaptar esos pesos se conoce como algoritmo de entrenamiento o aprendizaje. Los elementos citados permiten clasificar los diferentes tipos de redes encontradas



(a) Ejemplo de red neuronal.



(b) Ejemplo de una neurona.

Figura 2.3: Modelos de red neuronal y neurona.

en la literatura según los siguientes aspectos [18]:

Número y disposición de las neuronas. En una red de tipo *feedforward* las neuronas están organizadas en capas, y no hay conexiones recurrentes o realimentación de la salida sobre la entrada o entre neuronas internas. Existen, entonces capas directamente conectadas a la entrada del sistema, capas directamente conectadas a la salida, y capas ocultas o intermedias. En una red recurrente, por el contrario, existe realimentación entre neuronas lo que determina la aparición de una dinámica más compleja.

No-linealidad presente en cada neurona. La salida de cada neurona es, en general una función no lineal de la función de entrada.

Red de interconexión. Los valores de los pesos de la red de interconexión imponen la influencia que la salida de una neurona tiene sobre las otras. Esta influencia puede ser excitatoria, cuando tiende a aumentar el valor de la activación, o inhibitoria, cuando tiende a disminuirlo.

Algoritmo de entrenamiento. En los algoritmos supervisados, existe un tutor o salida deseada que especifica que valor debería tener la salida

de la red para cada posible entrada dentro de un conjunto de datos de entrenamiento. El error entre la salida actual de la red y la deseada es utilizada por el algoritmo para sintonizar el valor de los pesos de la red. En los algoritmos no supervisados, por el contrario, no existe un tutor y el objetivo del algoritmo de entrenamiento es detectar alguna regularidad, estructura o característica particular de los datos, de forma que la red se convierte en un detector de características (feature detector). Los algoritmos de entrenamiento son, generalmente, procesos de optimización, a menudo heurísticos y como tales sufren de problemas de convergencia y existencia de mínimos locales en la superficie de la búsqueda.

Comportamiento estático y dinámico. Una vez entrenadas, las redes que no poseen memoria se comportan como combinadores no lineales. Por el contrario, la existencia de elementos capaces de almacenar valores del estado pasado de la red la transforman en un sistema dinámico no lineal, lo que amplía significativamente sus capacidades de modelado y agrega complejidad a su comportamiento y tratamiento matemático.

2.2.3. Aprendizaje inductivo

Nuestra definición 6 de aprendizaje artificial expresa un proceso inductivo, donde dada una cantidad limitada de datos en forma de un conjunto de entrenamiento, tratamos de inducir a una función que se aproxima al proceso de etiquetado original sobre el universo de datos.
Es decir, generalizamos a partir de específicas instancias del conjunto de entrenamiento D a todo el universo de datos X. A este comportamiento lo llamamos aprendizaje inductivo. El corazón del aprendizaje inductivo cae en la suposición de que el conjunto de entrenamiento es una representación fiel del universo entero.
Esta precisión es formalizada en la siguiente hipótesis:

Hipótesis 1. *Aprendizaje inductivo [12].*
Cualquier función encontrada para aproximar la función objetivo más un conjunto suficientemente grande de ejemplos de entrenamiento también se aproximará a la función objetivo sobre los ejemplos no observados.

La pregunta en aprendizaje inductivo si un conjunto de entrenamiento es una buena representación de un universo de datos, no es una respuesta

clara. Es deseable, sin embargo, la construcción de un conjunto de entrenamiento representativo del universo de datos como sea posible, es decir, que es conveniente la construcción de un conjunto de entrenamiento que sea *suficientemente grande*.
Las sofisticadas técnicas de la teoría del muestreo estadístico pueden ser usadas para asegurar que el conjunto de entrenamiento sean *lo suficientemente grande*. Sin embargo, finalmente siempre habrá cierta incertidumbre en los objetos contenidos en nuestro conjunto de entrenamiento. Es importante estudiar las técnicas que nos ayudan a evaluar esta incertidumbre y, con ello, tener una capacidad de generalización y precisión esperada en nuestros modelos [12].

2.2.4. Maquinas de soporte vectorial

Las maquinas de soporte vectorial (en inglés *Support Vector Machines*) fueron introducidas por Vapnik a partir de sus trabajos sobre las teorías del aprendizaje estadístico, en los que acotaba el error de generalización en función de la complejidad del espacio de hipótesis. Los primeros trabajos datan de principios de los noventa. Desde entonces las maquinas de soporte vectorial han ganado un merecido reconocimiento gracias a los sólidos principios teóricos en los que se fundamenta su diseño y al estupendo rendimiento que ofrecen en una gran variedad de aplicaciones prácticas. Buena parte de su popularidad radica, precisamente, en el hecho de que las máquinas de soporte vectorial son capaces de producir buenos modelos para múltiples tipos de aplicaciones prácticas. Y aunque en su origen se diseñaron para resolver solamente problemas de clasificación binaria, en la actualidad su aplicación se ha extendido a tareas de regresión, multiclasificación, agrupamiento, y muy recientemente se empiezan a sentar las bases teóricas para resolver problemas en los que la salida puede ser aún más compleja y estructurada, como un árbol o un grafo.
¿Qué hace diferentes a las maquinas de soporte vectorial de otros métodos de aprendizaje? ¿Dónde reside su éxito? La mayor diferencia entre las máquinas de soporte vectorial y otros métodos tradicionales de aprendizaje es que las SVM no se centran en construir una hipótesis que cometa pocos errores, sino que lo que pretenden es producir predicciones en las que se pueda tener mucha confianza, aún a costa de cometer ciertos errores. Tradicionalmente, la mayoría de los algoritmos de aprendizaje se han centrado en lo primero, reducir al mínimo los errores cometidos por el modelo generado. Se basan

en lo que se denomina el principio de minimización del riesgo empírico. El enfoque de las máquinas de soporte vectorial es diferente, no buscaran reducir el riesgo empírico cometiendo pocos errores, sino que pretenden construir modelos confiables. Ese principio recibe el nombre de minimización del riesgo estructural. Es decir, las SVM buscan un modelo que estructuralmente tenga poco riesgo de cometer errores ante datos futuros.
Sin embargo, en su versión original para clasificación binaria, su planteamiento no es más complejo que el del perceptrón mostrado anteriormente. En su forma de resolverlo obviamente sí. Ambos son clasificadores lineales, es decir, producen el mismo tipo de modelos: una función lineal que podremos enriquecer (convertirla en no lineal) introduciendo funciones kernels. Esa es la primera idea de su diseño: se parte de un tipo de funciones sencillas y se busca una solución óptima. Después, se amplia el tipo de funciones que pueden aprenderse usando kernels, sin aumentar apenas la complejidad del proceso.
La diferencia entre el perceptrón y las maquinas de soporte vectorial está en la forma en la que cada sistema busca esa función lineal. Mientras el perceptrón trata en cada iteración de reducir el número de errores, las maquinas de soporte vectorial buscan el hiperplano que nos ofrezca menos riesgo desde un punto de vista estructural. ¿Cómo lo consiguen? Ahí es donde aparece uno de los elementos importantes de las máquinas de soporte vectorial, como es el concepto de margen. El objetivo de una máquina de soporte vectorial de clasificación es maximizar el margen de la solución, que en primera instancia podemos entender como producir el clasificador que separe en mayor medida las dos clases. El proceso de maximizar el margen nos llevara a un problema de optimización, que tendrá la ventaja de que nos calculará el hiperplano de mayor margen en un tiempo polinomial y sin la posibilidad de que el algoritmo se quede atrapado en un máximo local.

2.2.5. Agrupamiento

Clasificar casos u objetos en diferentes grupos es una necesidad en el mundo en el que vivimos donde hace falta poner orden y agrupar desde libros o sombreros, hasta estrellas, organismos, moléculas, o la información que tantas veces buscamos en internet. El análisis de grupos o agrupamiento es una colección de métodos estadísticos que permiten agrupar casos sobre los cuales se miden diferentes variables o características. Así, los casos que presenten características muy similares deberán quedar agrupados en conjuntos que

llamaremos grupos, estos grupos deberán ser hallados sin información previa y serán sugeridos únicamente por la propia esencia de los datos. También hay que comentar que en los últimos avances en el estudio de las secuencias del ADN y del genoma humano han sido muchas veces posibles gracias a novedosos algoritmos de análisis de grupos.

El agrupamiento (en inglés, clustering), también conocido como partición de conjuntos, es una metodología básica y ampliamente aplicada. Las áreas de aplicación incluyen: estadística, programación matemática y ciencias de la computación (reconocimiento de patrones, teoría de aprendizaje, procesamiento de imágenes, gráficos por computadora, etc.). El agrupamiento consiste principalmente en agrupar todos los objetos en conjuntos (clusters) mutuamente excluyentes para alcanzar el máximo o mínimo de una función objetivo [5].

Uno de los problemas fundamentales del análisis de grupos es que no existe una definición precisa de grupo (cluster). Ello ha dado lugar al desarrollo de una gran cantidad de métodos; así, podremos hablar de dos grandes bloques de métodos de agrupamiento: los jerárquicos o no jerárquicos o particionales. En los primeros, la pertenencia a un grupo en un nivel de la jerarquía condiciona la pertenencia a grupos de un nivel superior. Además, se dividen en aglomerativos o divisibles, según la jerarquía que sea construida agrupando casos o bien dividiendo secuencialmente los datos. Los métodos particionales obtienen una única partición de los datos mediante la optimización de una función adecuada. Estos métodos también son conocidos como métodos de optimización. Es importante mencionar que en esta tésis se usan los métodos heurísticos, en específico los algoritmos genéticos.

Los algoritmos heurísticos de agrupamiento pueden ser divididos en cuatro categorías: Agrupamiento en la estadística convencional, programación matemática, flujo en redes y algoritmos genéticos [5].

Modelo matemático de agrupamiento

El modelo matemático de agrupamiento de un número (m) de clusters es:

$$\begin{aligned} & [CA_m] & \\ & Max\ F(X) & \quad sujeta\ a \\ & \sum_j X_{ij} = 1, & \quad para\ todo\ i, \\ & \sum_j X_{jj} = m, & \\ & X_{ij} \leq X_{jj} & \quad para\ todo\ i,j, \\ & X_{ij} = \{0,1\} & \quad para\ todo\ i,j, \end{aligned}$$

donde $X_{ij} = 1$ indica que el i-ésimo objeto es asignado al j-ésimo cluster, $X_{ij} = 0$ en otro caso, $i,j = \{1,2,...,N\}$, N es el número de objetos, m es el número de clusters, $F(X)$ es la función objetivo.

Distancias y similaridades

Supongamos m casos recogidos en un conjunto que llamaremos Ω y denotaremos $\Omega = \{1,2,\cdots,m\}$. Teniendo en cuenta que nuestro objetivo principal es hallar grupos que contengan los casos similares, va a ser necesario medir las similitudes o bien las distancias que hay entre los casos.

Definición 7. *Distancia entre grupos*
Una distancia sobre un conjunto Ω es una función d:

$$\begin{aligned} d: \ \Omega \times \Omega & \mapsto \mathbb{R} \\ (i,j) & \mapsto d(i,j) = d_{i,j} \end{aligned}$$

tal que verifica las siguientes propiedades:

1. $d(i,j) \geq 0,\ \forall i,j \in \Omega$
2. $d(i,j) = 0,\ \forall i \in \Omega$
3. $d(i,j) = d(j,i),\ \forall i,j \in \Omega$

La primera de las propiedades dice que todas las distancias tienen que ser no negativas. La segunda propiedad dice que cada caso no puede distar de si mismo y la última de las propiedades establece la simetría. Es decir, que la distancia que puede haber de un caso i al otro caso j es la misma que del caso j al caso i. En general cuanto mayor sea la distancia $d(j,i)$, más diferentes entre si serán los casos i y j.
Como el número de casos m es finito, podemos ordenar las interdistancias en una matriz simétrica $m \times m$, que llamaremos matriz de distancia sobre Ω:

$$D = \begin{pmatrix} d_{11} & d_{12} & \cdots & d_{1m} \\ d_{21} & d_{22} & \cdots & d_{2m} \\ \vdots & \vdots & & \vdots \\ d_{m1} & d_{m2} & \cdots & d_{mm} \end{pmatrix}$$

El concepto dual a la distancia es la similaridad. En algunos casos puede ser más práctico calcular similaridades que distancias.

Definición 8. *Similaridad entre grupos*
Una similaridad sobre un conjunto Ω es una función s:

$$\begin{aligned} s: \ \Omega \times \Omega &\mapsto \mathbb{R} \\ (i,j) &\mapsto s(i,j) = s_{i,j} \end{aligned}$$

tal que:

1. $0 \leq s(i,j) \leq 1, \ \forall i,j \in \Omega$
2. $1 = s(i,i) \geq s(i,j), \ \forall i \in \Omega$
3. $s(i,j) = s(j,i), \ \forall i,j \in \Omega$

La primera propiedad dice que la similaridad debe ser no negativa y establece una escala. La segunda, que cada caso se parece a sí mismo más que a cualquier otro caso y la ultima establece la simetría. Y en cuanto a la interpretación se puede decir que cuanto mayor sea la similaridad $s(i,j)$, más parecidos serán entre sí los casos i y j.
De la misma manera que construímos la matriz de distancias podemos construir la matriz de similaridades sobre Ω,

$$S = \begin{pmatrix} s_{11} & s_{12} & \cdots & s_{1m} \\ s_{21} & s_{22} & \cdots & s_{2m} \\ \vdots & \vdots & & \vdots \\ s_{m1} & s_{m2} & \cdots & s_{mm} \end{pmatrix}$$

Una vez construida una matriz de similaridades S hay varias formas de pasar a una matriz de distancia sobre Ω y viceversa.
En general, sobre cada caso de Ω se habran medido n variables y por lo tanto, cada caso puede ser representado por un punto $x = \{x_1, \cdots, x_n\}$ de $\mathbb{R}^n$, de manera que cada x_i es el valor que toma la i-ésima variable X_i medida sobre el caso. Dependiendo de la naturaleza de las variables que se hayan considerado (variables continuas binarias o mixtas), se deben utilizar diferentes tipos de distancias o similaridades. Además el uso de una distancia (similaridad) u otra también depende de la naturaleza de los datos, es decir si provienen de un estudios genético, ecológico, industrial, etc. Así pues hay una variedad enorme de diferentes funciones de distancias y similaridades.
Distancias para variables continuas
Supongamos n variables continuas. A continuación se presentan algunos ejemplos de distancias estadísticas entre dos casos de Ω representados por los puntos $x = (x_1, \cdots, x_n)'$ e $y = (y_1, \cdots, y_n)'$.

Distancia euclideana,

$$\begin{aligned} d_E(x,y) &= \left[(x-y)'(x-y)\right]^{\frac{1}{2}} \\ &= \left[\sum_{i=1}^{n}(x_i - y_i)^2\right]^{\frac{1}{2}} \end{aligned}$$

Distancia de Minkowsky ($q \geq 1$),

$$d_M(x,y) = \left(\sum_{i=1}^{n}|x_i - y_i|^q\right)^{\frac{1}{q}}$$

Cuando $q = 2$ ésta se reduce a la distancia euclídea. Cuando $q = 1$, se obtiene la distancia también conocida como distancia ciudad o métrica de Manhattan.

Distancia valor absoluto,

$$d_{ABS}(x,y) = \sqrt{\sum_{i=1}^{n}|x_i - y_i|}$$

Obsérvese que esta distancia es una modificación de la de Minkowsky con $q = 2$.

Distancia de Mahalanobis, Esta distancia es muy utilizada cuando los datos provienen de una o varias poblaciones pero con matriz de varianzas-covariazas Σ, común.

$$d_{MH}(x,y) = \left[(x-y)'\Sigma^{-1}(x-y)\right]^{\frac{1}{2}}$$

Similaridades para variables binarias
Cuando todas las variables $X_1, \cdots, X_n$ medidas sobre los casos son binarias, es decir, solamente toman los valores 0 ó 1, es más fácil calcular las similaridades para luego transformar éstas en distancias. Habitualmente el valor 0 indica que la característica en estudio no está presente, mientras que el valor 1 indica la presencia de la característica. Consideremos los casos i y j de Ω representados como $x = (x_1, \cdots, x_n)'$ e $y = (y_1, \cdots, y_n)'$. Para calcular la similaridad entre ellos nos basaremos en la tabla 2.3. En esta tabla se resume el recuento de las coincidencias de los valores que han tomado las n variables en los dos casos. Es decir:

- los dos casos han tomado el valor 1 simultáneamente en a variables,
- los dos casos han tomado el valor 0 simultáneamente en d variables,
- el caso i ha tomado el valor 0 mientras que el caso j ha tomado el valor 1 en b variables,
- el caso i ha tomado el valor 1 mientras que el caso j ha tomado el valor 0 en c variables.

		Caso i		
		1	0	
Caso j	1	a	b	$a+b$
	0	c	d	$c+d$
		$a+c$	$b+d$	n

Tabla 2.3: Recuento de las coincidencias de n variables binarias definidas para dos casos i y j, con $n = a+b+c+d$.

Hay muchas maneras de definir las similaridades en base a las cantidades a, b, c y d, a continuación se definen dos de las más habituales.

Similaridad de Sokal-Michener,

$$S_{SM}(i,j) = \frac{a+d}{n}$$

Similaridad de Jaccard

$$S_J(i,j) = \frac{a}{a+b+c}$$

Similaridad para variables mixtas
Cuando las n variables consideradas son mixtas, es decir, cuando tenemos n_1 variables cuantitativas, n_2 variables binarias y n_3 variables cualitativas ($n = n_1 + n_2 + n_3$), es muy habitual utilizar la distancia de Gower. La distancia de Gower entre los casos i y j se calcula como $d_{ij} = (1 - S_{ij})^{\frac{1}{2}}$ con

$$S_{ij} = \frac{\sum_1^{n_1} \left(1 - \frac{|x_l - y_l|}{R_l}\right) + a + \alpha}{n_1 + (n_2 - d) + n_3}$$

donde:

- $x_1, \cdots, x_{n_1}$ e $y_1, \cdots, y_{n_1}$ representan los valores observados de las variables cuantitativas para los casos i y j
- R_l es el rango de la l-ésima variable cuantitativa.
- a y d son los mismos recuentos de las coincidencias que se han comentado en la tabla 2.3 para las n_2 variables binarias.
- α es el número de coincidencias entre las variables cualitativas.

2.2.6. Redes bayesianas

Las redes bayesianas modelan un fenómeno mediante un conjunto de variables y las relaciones de dependencia entre ellas. Dado este modelo, se puede hacer inferencia bayesiana; es decir, estimar la probabilidad posterior de las variables no conocidas, en base a las variables conocidas. Estos modelos pueden tener diversas aplicaciones, para clasificación, predicción, diagnóstico, etc. Además, pueden dar información interesante en cuanto a cómo se relacionan las variables del dominio, las cuales pueden ser interpretadas en

ocasiones como relaciones de causa-efecto.
Inicialmente, estos modelos eran construidos 'a mano' basados en un conocimiento experto, pero en los últimos años se han desarrollado diversas técnicas para aprender a partir de datos tanto la estructura como los parámetros asociados al modelo. También es posible el combinar conocimiento experto con los datos para aprender el modelo.

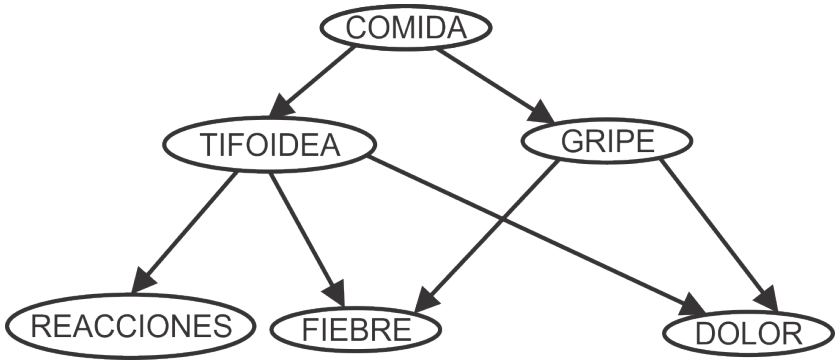


Figura 2.4: Ejemplo de red bayesiana.

Las redes bayesianas son una representación gráfica de dependencias para razonamiento probabilístico, en la cual los nodos representan variables aleatorias y los arcos representan relaciones de dependencia directa entre las variables. La figura 2.4 muestra un ejemplo hipotético de una red bayesiana (RB) que representa cierto conocimiento sobre medicina. En este caso, los nodos representan enfermedades, síntomas y factores que causan algunas enfermedades. La variable a la que apunta un arco es dependiente de la que está en el origen de éste, por ejemplo *fiebre* depende de *tifoidea* y *gripe* en la red de la figura 2.4. La topología o estructura de la red nos da información sobre las dependencias probabilísticas entre las variables. La red también representa las independencias condicionales de una variable (o conjunto de variables) dada(s) otra(s) variable(s).

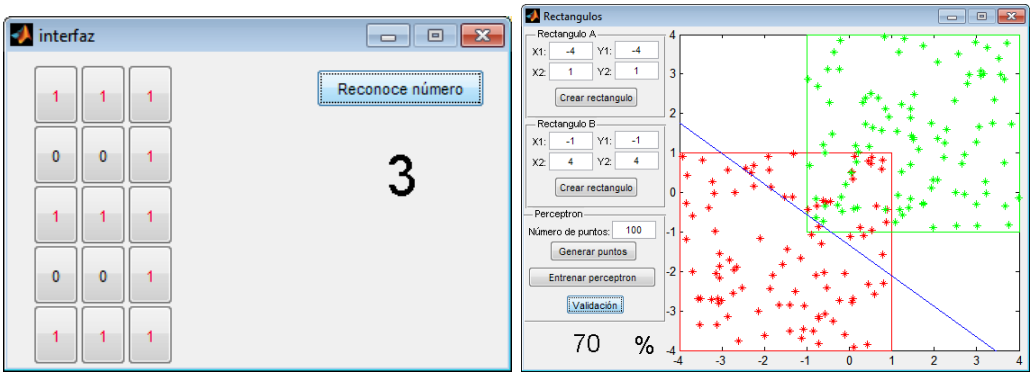
2.3. Estudio de las técnicas de aprendizaje artificial

Con la finalidad de comprender el estado del arte del área de aprendizaje artificial se procedió a analizar las técnicas mencionadas en la sección anterior

para analizar como abordar el análisis de calidad del aire. A continuación se describen de manera breve las técnicas estudiadas, así como las aplicaciones realizadas.

2.3.1. Redes neuronales artificiales

La arquitectura inicial a estudiar dentro de las redes neuronales es el perceptron, pues de este se parte para abordar arquitecturas mas detalladas, el perceptron es usado para tareas simples de reconocimiento de patrones, para entender esta arquitectura simple se programaron dos aplicaciones que muestran la clasificación que se puede hacer con el perceptron, en la figura 2.5 (inciso a) se muestra el reconocimiento de números , así como la clasificación de dos conjuntos de datos (inciso b).



(a) Clasificador de números. (b) Clasificador de conjuntos.

Figura 2.5: Ejemplos de clasificación con perceptron.

Una vez comprendido el tipo de aplicaciones que se pueden realizar con un perceptron simple, se procedió a estudiar el perceptron multicapa, en este caso se tomó el conjunto de datos de la flor de iris, con el propósito de crear un reconocedor de los tipos de la flor de iris con los datos de entrada, la figura 2.6 muestra el funcionamiento de la aplicación desarrollada.
Es importante mencionar que existe un gran número de topologías, es por ello que se estudiaron las memorias asociativas, estas técnicas se basan en la idea de que los seres humanos asociamos el conocimiento, es así como se estudio la memoria asociativa bidireccional y la de Hopfield, para no solo abordar la teoría elemental de redes neuronales, si no entender los beneficios

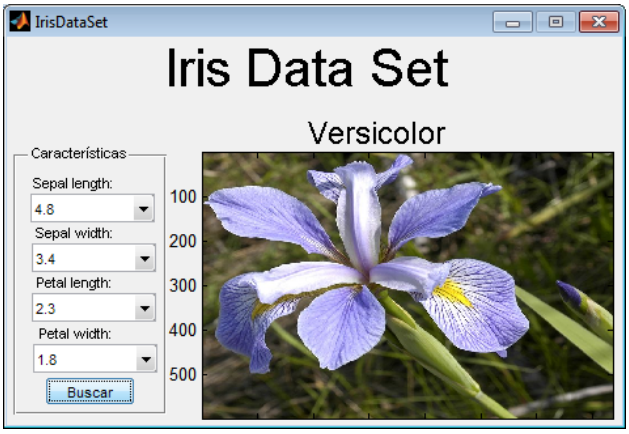


Figura 2.6: Reconocedor de la flor de iris con perceptron multicapa.

de las arquitecturas. La figura 2.7 muestra el reconocedor de números con el perceptron simple y las memorias asociativas antes mencionadas.

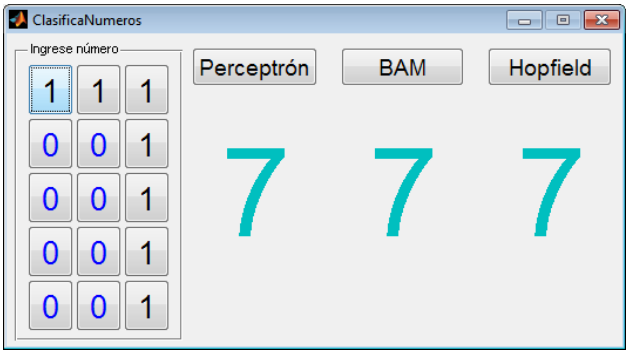


Figura 2.7: Reconocedor de números con perceptron, Hopfield y BAM.

2.3.2. Maquinas de soporte vectorial

Las maquinas de soporte vectorial son modelos de aprendizaje supervisado asociadas con algoritmos de aprendizaje, las cuales analizan los datos, para así reconocer patrones. Una maquina de soporte vectorial toma un conjunto de datos de entrada y predice para cada entrada dada cual de dos posibles clases es la salida haciendo un clasificador lineal binario no probabilístico. Dado un conjunto de ejemplos de entrenamiento que pertenecen a

una de dos categorías, una maquina de soporte vectorial construye un modelo que asigna ejemplos desconocidos a una categoría o a la otra. El modelo de una maquina de soporte vectorial es una representación de los ejemplos de entrenamiento como puntos en un espacio dimensional, donde se construye una separación que divide a las dos categorías. Los ejemplos desconocidos son entonces mapeados al espacio construido para así predecir a que categoría pertenece dicho ejemplo basado en el lado de la separación en que se encuentra.
Además de desempeñar la clasificación lineal, las maquinas de soporte vectorial pueden desempeñar de manera eficiente clasificación no lineal usando distintos tipos de kernel y mapeando las entradas en espacios de características de dimensiones mas grandes.
Con la finalidad de entender el funcionamiento básico de las maquinas de soporte vectorial se probo la clasificación de la flor de iris, además de que se agregó la posibilidad de usar al mismo tiempo el perceptron, para observar que resultados daba cada una de estas técnicas con una entrada dada y observar el tiempo que se tardan en reconocer cada técnica, donde se notó que la maquina de soporte vectorial tarda menos tiempo para clasificar una entrada. La figura 2.8 muestra dicha aplicación.

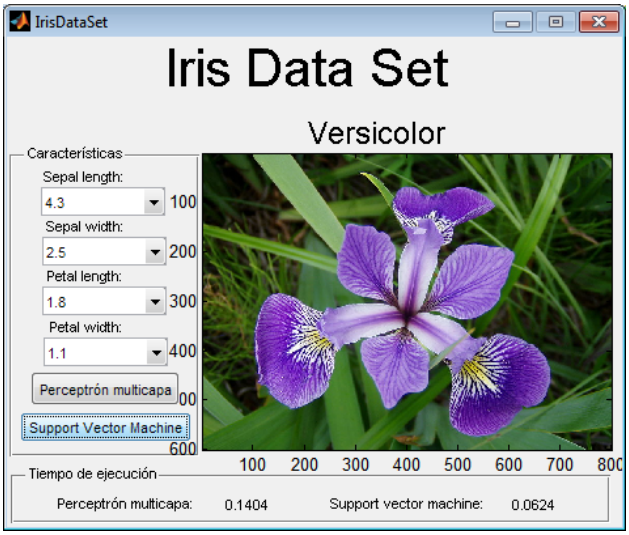


Figura 2.8: Reconocedor de la flor de iris con maquina de soporte vectorial.

2.4. Agrupamiento

Análisis de agrupamiento o agrupamiento es la tarea de asignar un conjunto de objetos en grupos (llamados clusters) de modo que los objetos en el mismo grupo son más similares entre ellos, que a los elementos en otros grupos.
El agrupamiento es una técnica explorativa común del análisis estadístico de datos y es usado en muchas áreas como aprendizaje artificial, reconocimiento de patrones, análisis de imágenes, recuperación de información y bioinformática.
Cuando se habla de análisis de grupos no se refiere a un algoritmo en específico, si no a la tarea a ser resuelta. Esta puede ser desempeñada por varios algoritmos que difieren de manera significante en su noción de como formar un grupo y como encontrarlo de manera eficiente.
A manera de estudio se procedió a realizar una aplicación para observar el funcionamiento de las técnicas clásicas de agrupamiento (BSAS, k-Means, Fuzzy c-Means). El conjunto de datos fue el de la flor de iris; el funcionamiento de la aplicación se puede observar en la figura 2.9.

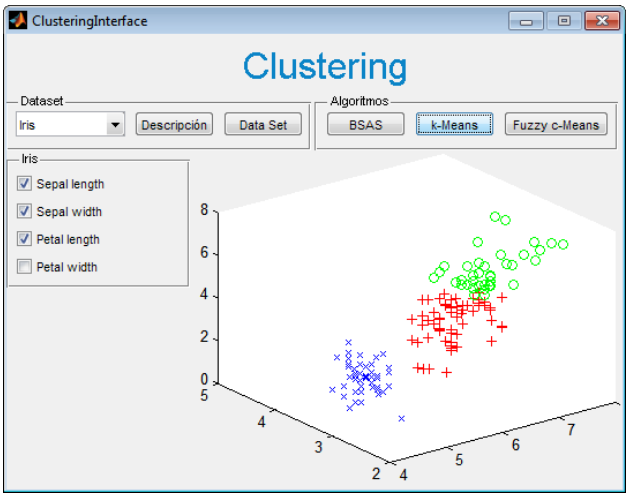


Figura 2.9: Agrupamiento de los datos de la flor de iris.

2.5. Conclusiones

En este capítulo, nos hemos enfocado al análisis de las técnicas de aprendizaje artificial. Hemos estudiado dentro del área de aprendizaje supervisado las técnicas de redes neuronales y maquinas de soporte vectorial. Por el lado del aprendizaje no supervisado se ha revisado la teoría involucrada al análisis de agrupamientos.
Es necesario identificar cual de estas técnicas se adapta en mejor medida al problema con el que vamos a tratar. Es así como decidimos usar las técnicas de agrupamiento pues, no contamos con una etiqueta o con grupos de pertenencia a una cierta característica, si no al contrario necesitamos identificar estos grupos y observar con que características cumplen.
Ahora bien, si bien los algoritmos de agrupamiento son una posible manera de atacar el problema, es importante mencionar que los datos con los que vamos a trabajar, son demasiados como para usar un algoritmo clásico de agrupamiento, es por ello que se decidió analizar los algoritmos de agrupamiento que se enfocan en la optimización de una función objetivo. Es por ello que en el siguiente capítulo se hablará de la técnica algoritmos genéticos, así como algoritmos que se valen de esta idea para buscar soluciones óptimas al problema de agrupamiento.

Capítulo 3

Algoritmos genéticos para agrupamiento

3.1. Computación evolutiva

La computación evolutiva o los algoritmos evolutivos engloban una serie de técnicas inspiradas biológicamente.
Para simular el proceso evolutivo en una computadora se requiere:

- Codificar las estructuras que se replicaran.
- Operaciones que afecten a los "individuos".
- Una función de aptitud.
- Un mecanismo de selección.

Paradigmas:

- Programación evolutiva.
- Estrategias evolutivas.
- Algoritmos genéticos.

Cada uno de estos paradigmas se originó de manera independiente y con motivaciones distintas.

3.1.1. Programación evolutiva

Esta enfatiza los nexos de comportamiento entre el caso de los padres e hijos, en vez de buscar emular operaciones genéticas específicas (como algoritmos genéticos). *Algoritmo:*

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

La programación evolutiva es una abstracción de la evolución al nivel de las especies, por lo que no se requiere el uso de un operador de recombinación (diferentes especies no se pueden cruzar entre si), igualmente, usa selección probabilista.
Aplicaciones: predicción, generalización, juegos, control automático, TSP, planeación de rutas, etc.

3.1.2. Estrategias evolutivas

Fueron desarrolladas en 1964 en Alemania para resolver problemas hidrodinámicos de alto grado de complejidad por un grupo de estudiantes de ingeniería encabezados por Ingo Recherberg.

Algoritmo

La versión original usaba un solo padre y con él se generaba un solo hijo. Este hijo se mantenía si era mejor que el padre, o de lo contrario se eliminaba. Se usa la selección extintiva, es decir, los peores individuos obtienen una probabilidad de ser seleccionados de cero.
En esta versión un individuo nuevo es generado usando:

$$x^{t-1} = x^t + N\left(0, \bar{\sigma}\right)$$

donde t se refiere a la *generación* (o iteración) en la que nos encontramos y $N\left(0, \bar{\sigma}\right)$ es un vector de números Gaussianos independientes con una medida de cero y desviación estándar $\bar{\sigma}$.
Es importante destacar que su selección es extintiva, es decir, los peores individuos obtienen una probabilidd de ser seleccionados de cero.

Estrategias evolutivas vs Programación evolutiva

La programación evolutiva usa normalmente selección estocástica, mientras que las estrategias evolutivas usan selección determinística.
Ambas técnicas operan a nivel fenotípico (es decir, no requieren codificación de las variables del problema).
la programación evolutiva es una abstracción de la evolución al nivel de las especies, por lo que no se requiere el uso de un operador de recombinación (diferentes especies no se pueden cruzar entre sí). En contraste las estrategias evolutivas son una abstracción de la evolución al nivel de un individuo, por lo que la recombinación es posible.

Aplicaciones de las estrategias evolutivas

Problemas de ruteo y redes, bioquímica, óptica, diseño e ingeniería, magnetismo.

3.1.3. Algoritmos genéticos

Los algoritmos genéticos son algoritmos de búsqueda de propósito general que usan principios inspirados en la genética natural de las poblaciones para desarrollar soluciones a problemas. La idea básica de los algoritmos genéticos es mantener una población de cromosomas que representan soluciones candidatas. Un cromosoma es compuesto de una serie de genes que representan variables de decisión o parámetros. Cada miembro de la población es evaluado y se le asigna una medida de su aptitud como una solución [5].
Los algoritmos genéticos (GAs, por sus siglas en ingles), se denominaron originalmente planes reproductivos genéticos,estos fueron propuestos inicialmente por John H. Holland [13]. son algoritmos de búsqueda que tienen las características de búsqueda estocástica, búsqueda multipunto, búsqueda directa y búsqueda paralela. Usan la función de aptitud para evaluar sus cromosomas, AGs pueden ser aplicados a varias funciones objetivo sin necesidad de información adicional en la búsqueda.
El algoritmo enfatiza la importancia de la cruza sexual (operador principal) sobre el de la mutación (operador secundario) y usa selección probabilista.
El algoritmo básico es el siguiente:

- Generar (aleatoriamente) una población inicial.

0	1	1	0	1	1	0	1	0	0	1	1	1	0	0	1
cadena 1				cadena 2				cadena 3				cadena 4			

Tabla 3.1: Cromosoma de un algoritmo genético.

- Calcular la aptitud de cada individuo.
- Seleccionar (probabilísticamente) en base a la aptitud.
- Aplicar operadores genéticos (cruza y mutación) para generar la siguiente población.
- Repetir hasta que cierta condición se cumpla.

En la tabla 3.1 se muestra un ejemplo de representación. A la cadena binaria se le llama "cromosoma". A cada posición de la cadena se le denomina"gene" y al valor dentro de esta posición se le llama "alelo".
Para poder aplicar el algoritmo genético se requiere de los siguientes componentes:

- Una representación de las soluciones potenciales del problema.
- Una forma de crear una población inicial de posibles soluciones (normalmente un proceso aleatorio).
- Una función de evaluación que juegue el papel del ambiente, clasificando las soluciones en términos de su "aptitud".
- Operadores genéticos que alteren la composición de los hijos que se producirán para las siguientes generaciones.
- Valores para los diferentes parámetros que utiliza el algoritmo genético (tamaño de la población, probabilidad de cruza, probabilidad de mutación, número máximo de generaciones, etc.)

Como representación, usaremos un alfabeto binario. La función de adaptación se debe diseñar para cada problema en forma específica.
Dado un cromosoma en particular, la función de adaptación le asigna un número real, que refleja el nivel de adaptación al problema de individuo representado por el cromosoma. Durante la fase reproductiva se seleccionan los individuos de la población para cruzarse y producir descendientes que constituirán, una vez mutados la siguiente generación de individuos. A continuación se presentan los tres operadores genéticos: selección, cruza y mutación:

Algoritmo 2 Ejemplo de un algoritmo genético simple

```
Inicio
  Generar una población inicial
  Calcular la función de evaluación de cada individuo
  Mientras NO TERMINADO
  Inicio // producir nueva generación
    Para tamaño-población / 2 hacer
    Inicio // ciclo reproductivo
      Seleccionar dos individuos de la anterior generación, para el
      cruce (probabilidad de selección proporcional a la función de
      evaluación del individuo).
      Cruzar con cierta probabilidad los dos individuos obteniendo
      dos descendientes.
      Mutar los dos descendientes con cierta probabilidad.
      Calcular la función de evaluación de los dos descendientes mu-
      tados.
      Insertar los dos descendientes mutados en la nueva generación.
    Fin
    Si la población ha convergido entonces
      terminado = cierto
  Fin
Fin
```

Operador de selección

Asigna las posibilidades de reproducción a cromosomas basados en su aptitud. Una ruleta Monte Carlo es usualmente empleada. Es decir, el cromosoma con mayor función de aptitud es el que tiene más posibilidades de ser seleccionado [11].
La selección de padres se efectúa al azar usando un procedimiento que favorezca a los individuos mejor adaptados, es decir, cada individuo tiene asignada una probabilidad de ser asignado que es proporcional a su función de aceptación (ruleta sesgada).
Los individuos bien adaptados se escogerán varias veces por generación (debido a su alta probabilidad de selección), mientras que los pobremente adaptados al problema, no se escogerán más que de vez en cuando.
Una vez seleccionados dos padres, sus cromosomas se combinan, utilizando

habitualmente los operadores de cruza y mutación.

Operador de cruza

Combina las características de las estructuras de dos cromosomas (padres) para formar dos descendientes. La forma más simple de hacer una cruza es cambiar un segmento de los padres. Cruza en un punto, cruza en dos puntos y cruza uniforme son comúnmente empleadas [11].
El operador de cruza, toma dos padres seleccionados y corta sus listas de cromosomas en una posición elegida al azar, para producir dos sublistas iniciales y dos sublistas finales. Posteriormente se intercambian las sublistas finales de cada individuo, produciéndose dos nuevos cromosomas completos. De este modo ambos descendientes heredan genes de cada uno de los padres. Esta forma de hacer la cruza se conoce como cruza en un punto.
El operador de cruza no se aplica a todos los pares de individuos que han sido seleccionados para emparejarse, si no que se aplica de manera aleatoria, normalmente con una probabilidad entre 0.5 y 1.0. En el caso de que el operador de cruza no se aplique, la descendencia se obtiene simplemente duplicando los padres.

Operador de mutación

Consiste en alterar uno o más genes de los descendientes en un porcentaje muy bajo, para evitar caer en óptimos locales. El descendiente resultante es evaluado y reinsertado en la población. Este proceso continua hasta que un criterio predeterminado (por ejemplo un número máximo de generaciones, un valor pequeño de mejora en la aptitud entre dos generaciones adyacentes o cierto grado de madurez) es alcanzado [11].
El operador de mutación se aplica a cada hijo de manera individual y consiste en la alteración aleatoria (normalmente con probabilidad pequeña) de cada componente del cromosoma.
Se puede pensar que el operador de cruza es más importante que el operador de mutación, ya que proporciona una exploración rápida del espacio de búsqueda, este último asegura que ningún punto del espacio de búsqueda tenga probabilidad cero de ser examinado, y es de mucha importancia para asegurar la convergencia de los algoritmos genéticos.

3.2. Problema

Una vez comprendido el estado del arte de cambio climático, así como en cuanto a Ciencias de la Computación se refiere , se ha procedido a partir de la teoría de descubrir conocimiento para construir una herramienta que nos permita visualizar y analizar el dominio del problema, de este modo poder desarrollar modelos, comparar los modelos creados con el fin de evaluar que tan efectivo es cada uno de ello y si estos nos pueden servir para identificar patrones en los datos que nos puedan beneficiar en la toma de decisiones.
Al no contar con un indicador que nos diga a que grupo pertenecen las mediciones de calidad del aire, no tenemos categorías conocidas, esto nos lleva a pensar en técnicas de clasificación no supervisadas. Es por ello que se han seleccionado las técnicas de agrupamiento, que se encuentran documentadas en el área de aprendizaje artificial. Un problema al trabajar con grandes cantidades de datos necesitaremos de métodos de aproximación que nos permitan agilizar el cómputo de los datos, es por ello que se agregó a la presente tesis el uso de algoritmos genéticos para de esta manera tratar de aproximar una solución de modo que se reduzca el tiempo de procesamiento.
En las siguientes secciones se presentan los algoritmos estudiados, asi como la evaluación de su eficiencia.

3.3. Algoritmos analizados

Comúnmente surgen situaciones en las cuales es deseable agrupar grandes cantidades de números de objetos, símbolos o personas en pequeños números de grupos mutuamente excluyentes, cada uno teniendo miembros lo mas parecidos posibles. Agrupar estos objetos facilita considerar y entender sus relaciones en grandes colecciones; esto usualmente incrementa la eficiencia del manejo. [19].
Es importante mencionar que existen diversas técnicas de agrupamiento y que cada una de ellas tiene fundamentos particulares que las hacen adecuadas para cierto tipo de problemas.
En la realización de esta tesis se presto atención a dos tipos de agrupamiento. El primero es el agrupamiento jerárquico de tipo aglomerativo y el segundo el agrupamiento con el uso de algoritmos genéticos.

3.3.1. Métodos jerárquicos

Estos métodos generan una sucesión de particiones, donde cada partición se obtiene uniendo o dividiendo grupos.
Dentro de los métodos jerárquicos se distinguen dos tipos:

1. Métodos aglomerativos
2. Métodos divisivos

Con los métodos *aglomerativos* los nuevos grupos se crean uniendo clusters. Es decir, en la partición inicial cada elemento forma un grupo. Empieza el proceso de aglomeración de manera que se van uniendo los grupos de dos en dos y finaliza cuando todos los casos forman un único grupo.
Con los métodos *divisivos*, al contrario, los nuevos grupos se crean dividiendo en clusters. Es decir, en la partición inicial todos los casos forman un único grupo. Se comienza por dividir los grupos (habitualmente en dos). El proceso puede seguir hasta que cada caso forme un único grupo.
La principal ventaja de los métodos aglomerativos es su rapidez. Por su parte los métodos divisivos tiene la ventaja de que parten de la información global que hay en los datos y que además el proceso de división no tiene por que seguir hasta que cada elemento forme un grupo. Sin embargo, estos métodos suelen ser muy lentos y en general aplicables solo para datos con pocos casos [18]. Esto hace que los métodos jerárquicos más habituales sean los aglomerativos.
A continuación se describe el primer algoritmo estudiado, el cual forma parte de los algoritmos jerárquicos de tipo aglomerativo.

Método aglomerativo de agrupamiento jerárquico

Agglomerative hierarchical clustering method (AHCM) en inglés, incluye una serie de uniones sucesivas. Inicialmente, hay tantos conjuntos (clusters) como objetos a clasificar. Estos grupos iniciales son unidos de acuerdo a su grado de mejora en los valores objetivo. Eventualmente, todos los subgrupos son unidos en un solo conjunto [19]. A continuación se muestran los pasos en AHCM para agrupar N objetos en un problema de maximización [5]:

Paso 0. Iniciar con N conjuntos, cada uno conteniendo un solo objeto, es decir $S_i^{(1)} = \{O_i\}, i = 1, \cdots, N$. Una matriz simetrica de incrementos de una función objetivo $MF = \{\Delta F_{ij},\ i, j = 1, \cdots, N, i \neq j\}$, donde

ΔF_{ij} representa el incremento del valor objetivo en caso de que el i-ésimo y el j-ésimo grupo, sean unidos en un solo grupo. Sea $k = 1$.

Paso 1. Si $\Delta Fvu = max\{\Delta F_{ij},\ i, j = 1, \cdots, N-k, i \neq j\}$ y $v > u$, entonces $S_u^{(k+1)} = S_u^{(k)} \cup S_v^{(k)}$,$S_1^{(k+1)} = S_1^{(k)}, \cdots, S_{u-1}^{(k+1)} = S_{u-1}^{(k)}, S_{u+1}^{(k+1)} = S_{u+1}^{(k)}, \cdots, S_v^{(k+1)} = S_{v+1}^{(k)}, \cdots, S_{N-v-1}^{(k+1)} = S_{N-v}^{(k)}$ y $S_{N-k}^{(k+1)} = \varnothing$. Calcular el valor objetivo $F(X)^{(k)}$ de la partición. Asignamos k=k+1.

Paso 2. Repetir el *Paso 1* hasta $k = N-1$. $F(X)^* = max\left\{F(X)^{(k)}, k = 1, \cdots, N-1\right\}$

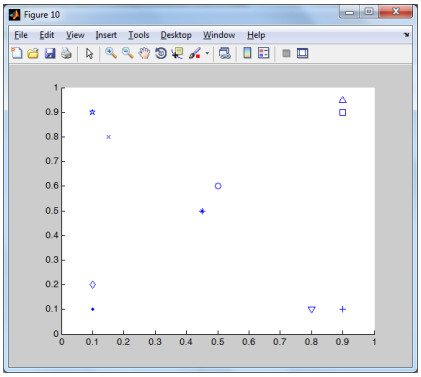
Un ejemplo de como se realiza el agrupamiento del método AHCM se muestra en la figura 3.1 donde se puede observar que inicialmente en el inciso a) todos los elementos se encuentran marcados con diferente forma, indicando que cada uno de ellos es un grupo, en el inciso b se observa el agrupamiento despues de cinco uniones, donde se encuentra la mejor solución al agrupamiento, finalmente el inciso c) muestra a todos los elementos en dos grupos, este es el resultado que se obtiene antes de que se agrupen todos los objetos en un único grupo.

3.3.2. Métodos con algoritmos genéticos

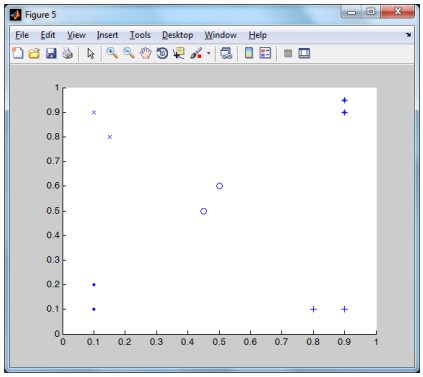
Los algoritmos de tipo jerárquico son una buena solución a problemas donde no existe una gran cantidad de datos, sin embargo al trabajar con miles de datos se intensifica el procesamiento que realizan los algoritmos jerárquicos. Debido a esto hemos optado por analizar métodos que se valen de la técnica de computo suave conocida como algoritmos genéticos, con la finalidad de encontrar la solución optima al problema de agrupamiento.
Otra ventaja que presentan los algoritmos genéticos es que buscan el número de grupos que mejore la solución a un problema dado. A continuación se describen los algoritmos estudiados.

Método simultaneo de agrupamiento

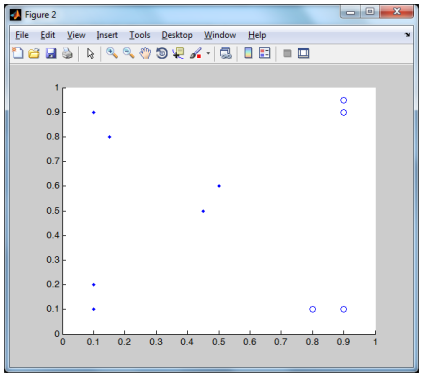
Simultaneously clustering method (SICM) en inglés. En un método de agrupamiento con algoritmos genéticos donde debe haber al menos dos objetos para formar un grupo; el máximo numero de grupos K, es igual a $[N/2]$ ([] es el signo gaussiano(Gauss sing)) y N es el número de objetos a agrupar. Entonces hay en total $N \times K$ variables de decisión, como lo muestra la tabla



(a) Agrupamiento inicial, los objetos pertenecen a grupos distintos.



(b) Solución que maximiza a la función objetivo con cinco grupos.



(c) Agrupamiento de dos grupos.

Figura 3.1: Ejemplo de agrupamiento de 10 objetos usando el algoritmo AHCM.

3.2. Si hay mas de dos variables con el valor de uno en la misma columna, esto quiere decir que estas forman un grupo.
Sin embargo si X_{ik} es codificado como un gen, la longitud del cromosoma será demasiado grande y podría agotar la memoria de la computadora. Además de que sería dificil manejar los cromosomas debido a la restricción de que estos solo pueden tener un solo gen igual a uno y los demas genes iguales a cero.
Para tratar con el problema, SICM usa una técnica de codificación y decodificación para remplazar cada fila de la matriz de la variable de decisión por una cadena de genes como se muestra en la tabla 3.3 [5].
En el caso de la tabla 3.3 un posible caso de tener cuatro grupos es que exis-

Objeto	Grupos					
	1	2	$\cdots$	k	$\cdots$	K
1	X_{11}	X_{12}		X_{1k}		X_{1K}
2	X_{21}	X_{22}		X_{2k}		X_{2K}
$\vdots$						
i	X_{i1}	X_{i2}		X_{ik}		X_{iK}
$\vdots$						
N	X_{N1}	X_{N2}		X_{Nk}		X_{NK}

Tabla 3.2: Relación entre N objetos y K grupos.

Genes	Núm. Decimal	Grupos
00	0	1
01	1	2
10	2	3
11	3	4

Tabla 3.3: Correspondenecia entre la cadena de genes, valor decimal y grupos.

tan 8 objetos, de este modo cada gen tendría dos valores asociados y cada cromosoma sería de longitug 16, un ejemplo de la correspondencia entre los objetos, la configuración del cromosoma y el grupo representado se muestra en la tabla 3.4.

Objeto	1		2		3		4		5		6		7		8	
Cromosoma	0	1	0	1	1	0	1	1	0	1	1	1	0	1	0	0
Grupo	2		2		3		4		2		4		2		1	

Tabla 3.4: Representación de grupos con el método SICM.

Una desventaja encontrada con el método SICM es que si el número de grupos a crear no es potencia de dos, existirán genes que representen grupos inexitentes, según lo especificado por el algoritmo SICM, lo cual ocasiona que al aplicar los operadores de mutación o cruza, se caiga en estos grupos y no se converga a una solución. Por ejempo si existen cinco grupos nos veremos forzados a tener genes de longitud tres para representar al grupo cinco, sin embargo tambin podemos representar a los grupos seis, siete y ocho al aplicar los operadores de los algoritmos genéticos. Así que debemos validar

que el gen sea valido, lo cual ocasiona mayor procesamiento.

Método de agrupamiento por etapas

Stepwise clusterig method (STCM) en inglés, resuelve de manera sucesiva el óptimo agrupamiento binario de un conjunto hasta que el valor objetivo no pueda ser perfeccionado. Un solo grupo inicial que contiene a todos los objetos es dividido en dos sub grupos de modo que la función objetivo es optimizada en este nuevo estado; a través de cada proceso de división binaria un grupo es dividido en dos subgrupos. Un grupo es llamado indivisible cuando este no puede mejorar la función objetivo cuando se agrupa de manera binaria. Este concepto es similar a un algoritmo de corte y poda. Cuando todos los grupos son indivisibles STCM ha obtenido el óptimo agrupamiento [5]. La figura 3.2 muestra un ejemplo de como se realiza la división del conjunto de objetos durante el procesamiento del algoritmo STCM.

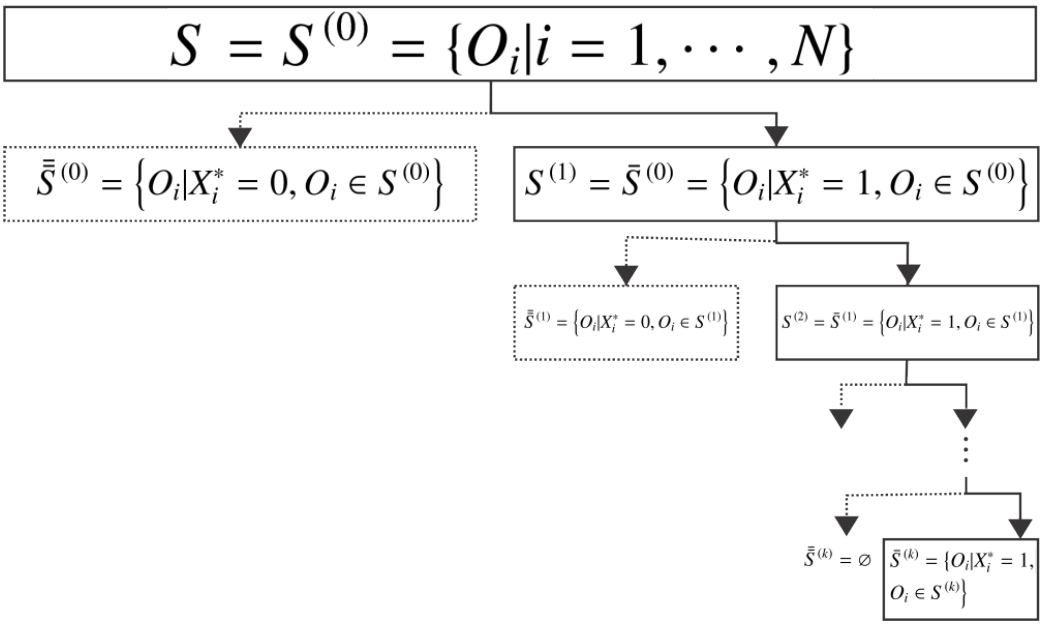


Figura 3.2: Procesamiento de algoritmo STCM.

Los pasos del algoritmo son:

Paso 0. Sea $S^{(0)}$ el grupo que contiene todos los objetos. El problema de dividir óptimamente $S^{(0)}$ en dos subgrupos $\bar{S}^{(0)}$ y $\bar{\bar{S}}^{(0)}$, puede ser resuelto con la siguiente programación matemática:

$$\begin{aligned} & MP^{(0)} \\ & Max\ F(X)^{(0)} \\ & \quad sujeta\ a\ X_i = \{0,1\} \quad i = 1, \cdots, \left|S^{(0)}\right|, \end{aligned}$$

donde $X_i = 1$ denota que el i-ésimo objeto de $S^{(0)}$ es agrupado en $\bar{S}^{(0)}$, $X_i = 0$ denota que el i-ésimo objeto de $S^{(0)}$ es agrupado en $\bar{\bar{S}}^{(0)}$; X_i es codificado como el gen de cromosomas (la longitud de cromosomas es $|S^{(0)}|$), entonces los algoritmos genéticos son empleados para resolver $MP^{(0)}$, maximizando $F(X)^{(0)}$ para obtener el óptimo agrupamiento binario: $\bar{S}^{(0)} = \left\{O_i | X_i^* = 1, O_i \in S^{(0)}\right\}$ y $\bar{\bar{S}}^{(0)} = \left\{O_i | X_i^* = 0, O_i \in S^{(0)}\right\}$. La tabla 3.5 muestra un ejemplo de cromosoma, donde se tienen 15 objetos, donde el valor de 1 indica la pertenencia a un grupo, y 0 la no pertenencia.

Cromosoma	1	1	0	1	0	0	1	1	0	0	1	0	0	0	0
Objeto	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15

Tabla 3.5: Representación de pertenencia a un grupo con algoritmo STCM.

Paso 1. Sea $S^{(1)} = \bar{S}^{(0)}$ y renumerando los objetos de $S^{(1)}$, se formula el óptimo agrupamiento binario de $S^{(1)}$ como $MP^{(1)}$, el cual también es resuelto por algoritmos genéticos. $F(X)^{(1)^*}$, es el valor objetivo del óptimo agrupamiento binario de $S^{(1)}$ bajo la suposición de que el otro grupo $\bar{\bar{S}}^{(0)}$ se mantiene sin cambio, el resultado del agrupamiento es: $\bar{S}^{(1)} = \left\{O_i | X_i^* = 1, O_i \in S^{(1)}\right\}$ y $\bar{\bar{S}}^{(1)} = \left\{O_i | X_i^* = 0, O_i \in S^{(1)}\right\}$. Tres grupos son formados $\bar{\bar{S}}^{(0)}$,$\bar{S}^{(1)}$ y $\bar{\bar{S}}^{(1)}$. $F(X)^{(1)^*}$ es el valor objetivo de estos tres grupos.

Paso 2. Sea $S^{(i)} = \bar{S}^{(i-1)}$ y resolver $MP^{(i)}$ mediante algoritmos genéticos.

Paso 3. Repetir el *Paso 2* hasta $\bar{\bar{S}}^{(k)} = \varnothing$, entonces este conjunto es indivisible; hay un total de $k+1$ grupos, es decir $\bar{\bar{S}}^{(0)}, \bar{\bar{S}}^{(1)}, \cdots, \bar{\bar{S}}^{(k-1)}$ y $\bar{S}^{(k)}$. $\bar{S}^{(k)}$ no puede ser mas dividido en los siguientes pasos. $F(X)^{(k)^*}$ es el óptimo valor objetivo de estos $k+1$ grupos.

Paso 4. Escoger uno de los restantes grupos para separarlo de manera binaria repitiendo los pasos *2* y *3* hasta que este sea indivisible.

Paso 5. Si todos los grupos son indivisibles, entonces termina. Los grupos formados son el resultado óptimo de SCTM, en otro caso, ir al *Paso 4*.

Método de agrupamiento basado en semillas

CSPM (Cluster seed points method, en inglés) emplea de manera inicial algoritmos genéticos para seleccionar las mas estables semillas para agrupar todos los objetos, después asigna el resto de objetos a cada grupo de acuerdo a su similitud con el grupo semilla o el grado de mejora de la función objetivo. El número de semillas representa el número de grupos y las características de estos grupos semilla determinan el resultado del agrupamiento.

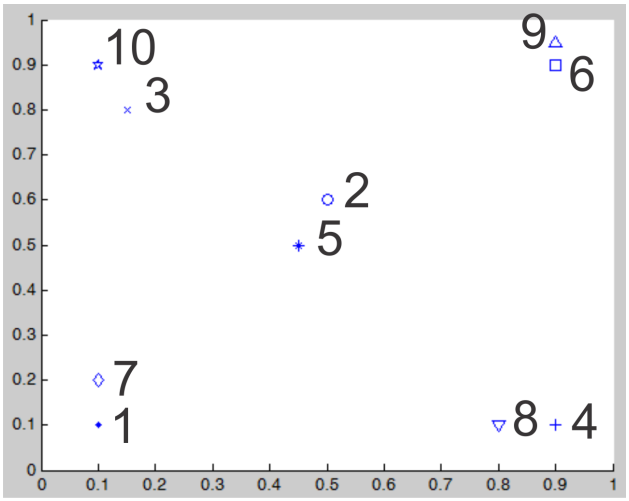


Figura 3.3: Ejemplo de objetos a ser procesados con el algoritmo CSPM.

Por ejemplo si tuvieramos los objetos mostrados en la figura 3.3, y se llegara a la configuración del cromosoma mostrado en la tabla 3.3, entonces los alelos con el valor de uno, representan las semillas que serán asignadas como el centro de cada grupo, es decir en la tabla 3.6 se tendrían cinco grupos.

cromosoma	1	1	1	1	0	1	0	0	0	0
grupo	1	2	3	4	NA	5	NA	NA	NA	NA

Tabla 3.6: Ejemplo de cromosoma del algoritmo CSPM. (NA = no asignado)

El algoritmo se presenta el algoritmo de asignación a continuación: [5].

Paso 0. Sea $k = 1$ y S un conjunto de todos los objetos, es decir, $S = O_1, \cdots, O_N$. CP_m es un conjunto de semillas, esto es $CP_m = C_1, \cdots, C_N$. NP es el conjunto de objetos que no son semilla $NP = S - CP_m$. $S_j^{(0)} = C_j, j = 1, \cdots, m$.

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

Paso 2. Sea $S_j^{(k)} = S_j^{(k-1)} \cup \{O_k\}$ y $k = k + 1$. Si $k < N - m + 1$ regresar al *Paso 1*, en otro caso terminar. Por ejemplo en

Una vez realizada la asignación, se obtendría un agrupamiento equivalente al mostrado en la figura 3.4.

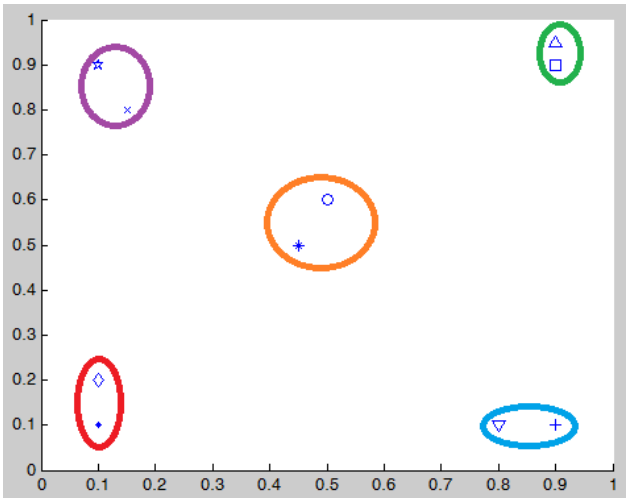


Figura 3.4: Ejemplo de grupos obtenidos con el cromosoma 3.6.

3.4. Evaluación

Una vez estudiadas las técnicas de agrupamiento es importante evaluar que tan eficaz es cada uno de los algoritmos estudiados, para así determinar que algoritmo provee mejores resultados o tarda menos tiempo de procesamiento. Existen criterios que nos permiten medir la calidad de una partición y nuestro objetivo será buscar la "mejor" partición según el criterio considerado. Para ello es importante introducir las siguientes matrices:

$$T = \frac{1}{m}\sum_{i=1}^{k}\left(\sum_{j=1}^{m_i}(x_{ij}-\bar{x})(x_{ij}-\bar{x})'\right), \tag{3.1}$$

$$W = \frac{1}{m-k}\sum_{i=1}^{k}\left(\sum_{j=1}^{m_i}\left(x_{ij}-\bar{x^i}\right)\left(x_{ij}-\bar{x^i}\right)'\right), \tag{3.2}$$

$$B = \frac{1}{m}\sum_{i=1}^{k} m_i\left(\bar{x^i}-\bar{x}\right)\left(\bar{x^i}-\bar{x}\right)', \tag{3.3}$$

donde x_{ij} es el vector de coordenadas del j-ésimo elemento del i-ésimo cluster C_i y donde $m_i = \#C_i$, $i = i, \cdots, k$.
Al checar en estas matrices vemos que en la matriz T determina las distancias entre los casos y el centro global, es decir, que en esta matriz se mide de alguna manera la variabilidad total que hay en los datos. En la matriz W se determinan las distancias entre los casos y el centro del grupo al que pertenecen. Luego esta matriz mide la variabilidad que hay dentro de los grupos. Finalmente, la matriz B determina las distancias entre los centros de los grupos y el centro global por lo que esta matriz mide la variabilidad entre los grupos. Además es conocido que se cumple la igualdad [18]:

$$T = W + B$$

Esta igualdad se puede traducir como:

$$\text{variabilidad total} = \begin{array}{c}\text{variabilidad}\\ \text{dentro}\\ \text{de los grupos}\end{array} + \begin{array}{c}\text{variabilidad}\\ \text{entre}\\ \text{los grupos}\end{array}$$

Para que una partición sea buena y contenga grupos con los casos más parecidos entre sí pero que no se parezcan a los casos de otros grupos, se necesitará que la variabilidad dentro de los grupos sea pequeña, o, dicho de otra

manera, que la variabilidad entre los grupos sea grande. Por eso, los criterios que se utilizan para encontrar dichas particiones buscarán de alguna manera hacer 'pequeña' la matriz W o, equivalentemente, hacer 'grande' la matriz B. Los criterios más habituales que se utilizan para hallar estas particiones son:

1. Minimizar la traza de W
2. Minimizar el determinante W
3. Minimizar el cociente de $\frac{det(W)}{det(B)}$
4. Maximizar la traza de $W^{-1}B$

Una vez elegido el criterio, se debe fijar el número de clusters; una técnica común es buscar el número de grupos k que optimice la partición entre todas las posibles. Esta no es tarea nada fácil ya que, dados m casos, el número de diferentes particiones en k clusters que se pueden hacer es del orden $\frac{k^m}{k!}$. Esto significa que si por ejemplo, $m = 50$ y $k = 2$, el número de posibles particiones es superior a 10^{14} y por tanto hay que buscar algoritmos que permitan encontrar la partición óptima sin tener que buscar entre todas las posibles.

3.4.1. Cálculo de la efectividad

En [5] los autores muestran como evaluar la efectividad y eficiencia de los algoritmos de agrupamiento, hemos reproducido este estudio, para corroborar los resultados presentados en el artículo; así como comparar otras propuestas y tener una base que nos garantice el buen funcionamiento de nuestros algoritmos. En [5] se propone un experimento sencillo donde se crean 200 objetos dos dimensionales. Todos los objetos son uniformemente distribuidos dentro de un cuadrado cuyas coordenadas de sus vértices son $(0, 0)$,$(0, 1)$,$(1, 0)$ y $(1, 1)$. Las pruebas se realizan con pocos objetos (10-50 objetos) para evaluar las técnicas en problemas pequeños, posteriormente se usan todos los objetos para probar con problemas de un número medio a grande (50-200 objetos). Generalmente se trata de minimizar la suma de los errores al cuadrado como función objetivo en los problemas de agrupamiento. Sin embargo es solo aplicable a problemas donde el número de grupos es especificado. Empleando esta función objetivo para resolver el número óptimo de grupos resultará en

N grupos tal que la suma de los errores al cuadrado es 0. En el estudio propuesto por [5] se propone como función objetivo la fórmula en un sentido ANOVA, con el fin de determinar simultáneamente el número óptimo de grupos y el resultado óptimo del agrupamiento.
La fórmula estadística en un sentido ANOVA F-Test es:

$$F = \frac{varianza\ determinada}{varianza\ indeterminada}$$

o

$$F = \frac{variabilidad\ entre\ los\ grupos}{variabilidad\ dentro\ de\ los\ grupos}$$

La varianza determinada o variabilidad entre los grupos es:

$$\sum_i n_i \frac{\left(\bar{Y}_i - \bar{Y}\right)^2}{(K-1)}$$

Donde $\bar{Y}_i$ denota la media del i-ésimo grupo. n_i es el número de observaciones en el i-ésimo grupo y $\bar{Y}$ denota la media total de los datos. La varianza indeterminada o variabilidad dentro de los grupos es:

$$\sum_{ij} \frac{\left(Y_{ij} - \bar{Y}_i\right)^2}{(N-K)}$$

Donde Y_{ij} es la j-ésima observación en la i-ésima salida de K grupos y N es el total de la muestra.
Es decir:

$$\begin{aligned} F &= \frac{\sum_i n_i \frac{\left(\bar{Y}_i - \bar{Y}\right)^2}{(K-1)}}{\sum_{ij} \frac{\left(Y_{ij} - \bar{Y}_i\right)^2}{(N-K)}} && (3.4) \\ &= \frac{\frac{1}{K-1} \sum_i n_i \left(\bar{Y}_i - \bar{Y}\right)^2}{\frac{1}{N-K} \sum_{ij} \left(Y_{ij} - \bar{Y}_i\right)^2} && (3.5) \\ &= \frac{(N-K)}{(K-1)} \times \frac{\sum_i n_i \left(\bar{Y}_i - \bar{Y}\right)^2}{\sum_{ij} \left(Y_{ij} - \bar{Y}_i\right)^2} && (3.6) \end{aligned}$$

Podemos observar que 3.6 coincide con la fórmula empleada en [5]

Ejemplo ANOVA en un sentido

Considere un experimento para estudiar el efecto de tres diferentes niveles de un factor en una respuesta (ej: tres niveles de un fertilizante para el crecimiento de plantas). Si tenemos seis observaciones por cada nivel, podríamos escribir el resultado del experimento en la siguiente tabla, donde a_1, a_2 y a_3 son los tres niveles del factor a ser estudiando.

a_1	a_2	a_3
6	8	13
8	12	9
4	9	11
5	11	8
3	6	7
4	8	12

Tabla 3.7: Valores de los tres niveles de un fertilizante.

La hipótesis nula H_0 para la prueba total de este experimento deberá ser que los tres niveles del factor producen la misma respuesta en promedio. Para calcular la razón de F.
Paso 1: Calculamos la media de cada grupo:

$$\begin{aligned}
\bar{Y_1} &= \frac{1}{6}\sum Y_{1i} = \frac{6+8+4+5+3+4}{6} = 5 \\
\bar{Y_2} &= \frac{1}{6}\sum Y_{2i} = \frac{8+12+9+11+6+8}{6} = 9 \\
\bar{Y_3} &= \frac{1}{6}\sum Y_{3i} = \frac{13+9+11+8+7+12}{6} = 10
\end{aligned}$$

Paso 2: Calcular la media total:

$$\bar{Y} = \frac{\sum_i \bar{Y_i}}{a} = \frac{\bar{Y_1}+\bar{Y_2}+\bar{Y_3}}{a} = \frac{5+9+10}{3} = 8$$

Donde a es el número de grupos.
Paso 3: Calcular la suma de cuadrados entre los grupos.

$$\begin{aligned}
S_B &= n\left(\bar{Y_1}-\bar{Y}\right)^2 + n\left(\bar{Y_2}-\bar{Y}\right)^2 + n\left(\bar{Y_3}-\bar{Y}\right)^2 \\
&= 6\left(5-8\right) + 6\left(9-8\right) + 6\left(10-8\right) = 84
\end{aligned}$$

Donde n es el número de objetos por grupo.
Los grados de libertad dentro del grupo es igual al número de grupos menos uno.

$$f_b = 3 - 1 = 2$$

Así, el valor de la media entre los grupos es:

$$MS_B = \frac{84}{2} = 42$$

Paso 4: Calcular la suma de cuadrados dentro de los grupos. Iniciando por centrar los datos en cada grupo, tabla 3.8.

a_1	a_2	a_3
6-5=1	8-9=-1	13-10=3
8-5=3	12-9=3	9-10=-1
4-5=1	9-9=0	11-10=1
5-5=0	11-9=2	8-10=-2
3-5=-2	6-9=-3	7-10=-3
4-5=-1	8-9=-1	12-10=2

Tabla 3.8: Calculo de la diferencia con el centro del grupo.

La suma de los cuadrados dentro del grupo es la suma de los cuadrados de todos los valores en la tabla 3.8.

$$S_W = 1+9+1+0+4+1+1+9+0+4+9+1+9+1+1+4+9+4 = 68$$

Los grados de libertad dentro del grupo son:

$$f_W = a\,(n-1) = 3\,(6-1) = 15$$

Así el valor de la media dentro del grupo es:

$$MS_W = \frac{S_W}{f_W} = \frac{68}{15} \approx 4.5$$

Paso 5: La razón F es:

$$F = \frac{MS_B}{MS_W} \approx \frac{42}{4.5} \approx 9.3$$

3.4.2. Prueba de Hipótesis

La prueba de hipótesis en el análisis de varianza de un sentido, nos sirve para comprobar si dos resultados son iguales con un cierto porcentaje de significancia o si estos no lo son.
En la prueba de hipótesis se comienza proponiendo una hipótesis tentativa acerca de un parámetro poblacional. Esta hipótesis tentativa se llama hipótesis nula y se representa con H_0. A continuación se define otra hipótesis, llamada hipótesis alternativa que es la opuesta de lo que se afirma en la hipótesis nula. La hipótesis alternativa se representa con H_a. El procedimiento para probar una hipótesis comprende el uso de datos de una muestra para probar las dos aseveraciones representadas por H_0 y H_a.

Establecimiento de la hipótesis nula y alternativa: En algunas aplicaciones puede no ser obvio como se deben formular las hipótesis nula y alternativa. Se debe tener cuidado para asegurar que las hipótesis estén bien estructuradas y que la conclusión de la prueba de hipótesis proporciona la información que desea el investigador o quien toma las decisiones. A continuación se describen tres tipos de situaciones en los que se emplean normalmente los procedimientos de pruebas de hipótesis.

Prueba de hipótesis de investigación: En estudios de investigación se deben formular la hipótesis nula y alternativa de tal modo que el rechazo de H_0 respalde la conclusión y la acción que se propone. En consecuencia, la hipótesis investigada debe expresarse como la hipótesis alternativa.

Prueba de la validez de una afirmación: En este tipo de prueba de hipótesis se supone, por lo general, que la afirmación del fabricante es verdadera, a menos que la evidencia parezca demostrar lo contrario. Si los resultados de la muestra indican que H_0 no se puede rechazar, no se puede dudar de la afirmación. Sin embargo, si estos resultados indican que se puede rechazar H_0, se afirmará la inferencia de que H_a es verdadera.
En cualquier caso donde intervenga la prueba de la validez de una afirmación, la hipótesis nula se suele basar en que la afirmación es verdadera. Entonces, se formula la hipótesis alternativa de tal modo que el rechazo de H_0 proporciona evidencia estadística de que lo afirmado es incorrecto. Se debe pensar en tomar acciones para corregir la afirmación, siempre que se rechace H_0 [2].

Prueba en caso de toma de decisiones: En la prueba de hipótesis de investigación o validez de una afirmación, se toman acciones si se rechaza H_0. Sin embargo, en muchos casos se deben emprender acciones tanto cuando no se puede rechazar H_0, como cuando si se puede rechazar. En general, estos casos se dan cuando quien toma las decisiones debe elegir entre dos cursos alternativos de acción, uno asociado con la hipótesis nula y otro con la hipótesis alternativa.

3.5. Resultados de las pruebas

Como se mencionó en la sección previa, se han realizado pruebas para evaluar la efectividad de los cuatro métodos estudiados, se ha usado la fórmula ANOVA para determinar la máxima razón de la variabilidad dentro de los grupos sobre la variabilidad entre los grupos. Además se han usado pruebas de hipótesis, con la hipótesis nula de que los valores obtenidos entre dos métodos son iguales con un 5 % de significancia.
La tabla 3.9 muestra que los valores encontrados por SICM son inferiores a los valores de AHCM mostrando la inefectividad de SICM, podemos notar que para N=10, 20 30 y 50 los valores encontrados por STCM son inferiores a los encontrados por AHCM, pero para N= 40 STCM muestra ser más efectivo que AHCM. Los valores resueltos por CSPM para N= 10, 20 y 30 no son significantemente diferentes sin embargo se puede observar que para N= 10 y 50 los resultados muestran ser más efectivos que los encontrados por AHCM. De este análisis podemos decir que CSPM es significantemente más efectivo que AHCM, seguidos por STCM y SICM es el menos efectivo debido a su gran cromosoma y a que la búsqueda de la función objetivo no se guía de un algoritmo de asignación o de evaluar que tan eficiente es el siguiente agrupamiento como es el caso de los demás algoritmos.
Una vez reproducidos los resultados mencionados, podemos mencionar que los resultados no difieren con [5], existen algunas diferencias poco significativas, pero esto se debe a que los objetos de prueba son generados de manera aleatoria, es decir estos resultados varian ya que no se usa el mismo conjunto datos. Con respecto a la prueba de hipótesis podemos decir que los valores obtenidos por los algoritmos STCM y CSPM son iguales con un 5 % de significancia.
En la tabla 3.10 se puede observar que a pesar de que CSPM es el método mas efectivo, también ocupa mayor tiempo de procesamiento comparado con

Número de objetos	AHCM F	SICM		STCM		CSPM	
		$\bar{F}_1$ ($\bar{F}_1/F$)	δF_1 (Z_1)	$\bar{F}_2$ ($\bar{F}_2/F$)	δF_2 (Z_2)	$\bar{F}_3$ ($\bar{F}_3/F$)	δF_3 (Z_3)
10	7.14	5.00	1.30	4.11	0.05	8.79	0.29
		(0.70)	(-8.97)	(0.57)	(-309.11)	(1.23)	(30.77)
20	33.74	4.15	1.16	24.88	2.13	33.22	1.04
		(0.12)	(-139.31)	(0.73)	(-22.72)	(0.98)	(-2.71)
30	40.55	3.58	0.87	33.51	6.15	40.49	2.87
		(0.08)	(-232.49)	(0.82)	(-6.26)	(0.99)	(-0.11)
40	31.20	2.66	0.51	37.50	9.26	55.74	6.64
		(0.08)	(-302.71)	(1.20)	(3.72)	(1.78)	(20.23)
50	47.08	2.20	0.44	32.83	4.48	54.04	4.32
		(0.04)	(-550.81)	(0.69)	(-17.38)	(1.14)	8.81

Tabla 3.9: Efectividad de los cuatro métodos estudiados ($N \leq 50$)

Número de objetos	Tiempo (segundos)			
	AHCM	SICM	STCM	CSPM
10	0.01	0.99	0.67	0.76
20	0.08	5.73	1.64	4.11
30	0.34	16.42	3.69	16.18
40	0.98	24.63	6.15	34.87
50	2.41	42.59	8.99	64.05

Tabla 3.10: Tiempo de procesamiento de los métodos estudiados ($N \leq 50$)

el resto de los algoritmos, por su parte STCM es el método con menor tiempo de procesamiento de los tres agrupamientos con algoritmos genéticos.

3.6. Conclusión

En el capítulo anterior se habló de los que es el área de aprendizaje artificial, una de las técnicas documentadas en esta área es el agrupamiento, esta técnica se utiliza cuando no se conoce a que clase o grupo pertenecen los datos que se van a procesar. Al mismo tiempo se habló del área de cómputo suave, un área donde se recurre a procesos bio-inspirados, tal es el caso de los algoritmos genéticos.

En este capítulo se retomaron ambos conceptos (algoritmos genéticos y agrupamiento), ya que en la presente tesis se ha puesto especial atención en algoritmos de agrupamiento que se apoyan en los algoritmos genéticos para mejorar la solución encontrada por algoritmos clásicos de agrupamiento por ejemplo los algoritmos jerárquicos de tipo aglomerativo.
Se ha analizado un algoritmo jerárquico y tres algoritmos genéticos para agrupamiento, la idea de tomar el algoritmo de tipo jerárquico se debe a que es importante tener una comparación para validar que tan efectiva es la solución encontrada por los algoritmos genéticos. Es así como se uso el análisis de varianza (ANOVA) para determinar la solución que maximiza la razón de la variabilidad entre los grupos sobre la variabilidad dentro de los grupos. Además se han usado pruebas de hipótesis documentadas en la teoría de estadística para verificar si las soluciones encontradas por dos métodos son iguales con un 5 % de significancia o no lo son.
Finalmente se realizó una prueba para evaluar la efectividad de los métodos documentados en esta capítulo, mostrando que el algoritmo más efectivo es CSPM, seguido por el algoritmo jerárquico AHCM mostrando que es posible encontrar soluciones más efectivas con la ayuda de los algoritmos genéticos, el algoritmo STCM provee soluciones similares a AHCM pero no presenta una gran mejora. El crear un espacio muy amplio de soluciones como es el caso del algoritmo SICM nos puede llevar a encontrar soluciones deficientes ya que este algoritmo mostro ser el menos efectivo en las pruebas realizadas además de que requiere mucho tiempo de procesamiento. Además se observó que el algoritmo AHCM provee resultados similares con los algoritmos CSPM y STCM con un 5 % de significancia.

Capítulo 4

Algoritmo propuesto y resultados experimentales

4.1. Algoritmo híbrido para agrupamiento

En el capítulo anterior se presentaron los algoritmos documentados en el área de agrupamiento que utilizan los algoritmos genéticos con la finalidad de hacer un análisis de la efectividad de estos comparados con un algoritmo de tipo jerárquico. Esta capítulo tiene como finalidad proponer un algoritmo híbrido que aproveche las bondades de los métodos estudiados para evaluar sí este provee mejores soluciones al problema de maximizar la razón de la variabilidad entre grupos sobre la variabilidad entre los grupos.

4.1.1. Descripción

En la sección del cálculo de la efectividad del capítulo anterior se pudo observar que el método más efectivo es el CSPM, sin embargo el algoritmo STCM es el algoritmo mas rápido de los algoritmos geneticos analizados. Es por ello que se ha decidido modificar el algoritmo STCM ya que consideramos que la creación de individuos por parte del algoritmo STCM es deficiente ya que contiene un amplio espacio de soluciones. Es por ello que se ha optado por usar el algoritmo STCM para cambiar la forma en que el algoritmo genético crea a los individuos de la población y usar el procedimiento de asignación que usa el algoritmo CSPM con la finalidad de observar si al crearse dos semillas y aplicar el algoritmo de asignación, ayuda al algoritmo STCM a obtener mejores resultados.

Método de agrupamiento por etapas con algoritmo de asignación

Este algoritmo retoma la idea presentada por el algoritmos STCM, la cual es resolver de manera sucesiva el óptimo agrupamiento binario de un conjunto hasta que el valor objetivo no puede ser perfeccionado. Un solo grupo inicial que contiene a todos los objetos es dividido en dos subgrupos de modo que la función objetivo es optimizada en este nuevo estado, el cambio a este algoritmo es que se seleccionan dos semillas aleatorias y se aplica un algoritmo de asignación (idea tomada del algoritmo CSPM) para aplicar entonces algoritmos genéticos y observar si esta asignación produce una mejora que supere el desempeño de STCM y CSPM. A través de cada proceso de división binaria un grupo es dividido en dos subgrupos y se vuelve a ejecutar el procedimiento de asignación. Un grupo es llamado indivisible cuando este no puede mejorar la función objetivo cuando se agrupa de manera binaria. Una vez que todos los grupos son indivisibles se ha obtenido el óptimo agrupamiento.
Los pasos del algoritmo son:

Paso 0. Sea $S^{(0)}$ el grupo que contiene todos los objetos. El problema de dividir óptimamente $S^{(0)}$ en dos subgrupos $\bar{S}^{(0)}$ y $\bar{\bar{S}}^{(0)}$, puede ser resuelto con la siguiente programación matemática:

$$\begin{aligned} &MP^{(0)} \\ &Max\ F\left(X\right)^{(0)} \\ &\quad sujeta\ a\ X_i = \{0,1\} \quad i = 1, \cdots, \left|S^{(0)}\right|, \end{aligned}$$

donde $X_i = 1$ denota que el i-ésimo objeto de $S^{(0)}$ es agrupado en $\bar{S}^{(0)}$, $X_i = 0$ denota que el i-ésimo objeto de $S^{(0)}$ es agrupado en $\bar{\bar{S}}^{(0)}$; X_i es codificado como el gen de cromosomas (la longitud de cromosomas es $|S^{(0)}|$); para realizar la creación de la cadena de cromosomas este algoritmo crea dos semillas y se asignan los genes restantes a una de las dos semillas de acuerdo a su cercania con cada una de las semillas. Entonces los algoritmos genéticos son empleados para resolver $MP^{(0)}$, maximizando $F\left(X\right)^{(0)}$ para obtener el óptimo agrupamiento binario: $\bar{S}^{(0)} = \left\{O_i | X_i^* = 1, O_i \in S^{(0)}\right\}$ y $\bar{\bar{S}}^{(0)} = \left\{O_i | X_i^* = 0, O_i \in S^{(0)}\right\}$.

Paso 1. Sea $S^{(1)} = \bar{S}^{(0)}$ y renumerando los objetos de $S^{(1)}$, se formula el óptimo agrupamiento binario de $S^{(1)}$ como $MP^{(1)}$, el cual también es resuelto por algoritmos genéticos. $F\left(X\right)^{(1)^*}$, es el valor objetivo del

óptimo agrupamiento binario de $S^{(1)}$ bajo la suposición de que el otro grupo $\bar{\bar{S}}^{(0)}$ se mantiene sin cambio, el resultado del agrupamiento es: $\bar{S}^{(1)} = \left\{O_i | X_i^* = 1, O_i \in S^{(1)}\right\}$ y $\bar{\bar{S}}^{(1)} = \left\{O_i | X_i^* = 0, O_i \in S^{(1)}\right\}$. Tres grupos son formados $\bar{\bar{S}}^{(0)}$,$\bar{S}^{(1)}$ y $\bar{\bar{S}}^{(1)}$. $F\left(X\right)^{(1)^*}$ es el valor objetivo de estos tres grupos.

Paso 2. Sea $S^{(i)} = \bar{S}^{(i-1)}$, aplicar el algoritmo de asignación y resolver $MP^{(i)}$ mediante algoritmos genéticos.

Paso 3. Repetir el *Paso 2* hasta $\bar{\bar{S}}^{(k)} = \varnothing$, entonces este conjunto es indivisible; hay un total de $k+1$ grupos, es decir $\bar{\bar{S}}^{(0)}, \bar{\bar{S}}^{(1)}, \cdots, \bar{\bar{S}}^{(k-1)}$ y $\bar{S}^{(k)}$. $\bar{S}^{(k)}$ no puede ser mas dividido en los siguientes pasos. $F\left(X\right)^{(k)^*}$ es el óptimo valor objetivo de estos $k+1$ grupos.

Paso 4. Escoger uno de los restantes grupos para separarlo de manera binaria repitiendo los pasos *2* y *3* hasta que este sea indivisible.

Paso 5. Si todos los grupos son indivisibles, entonces termina. Los grupos formados son el resultado óptimo, en otro caso, ir al *Paso 4*.

cromosoma	0	0	1	0	0	1	0	0	0	0
representación	NA	NA	grupo 1	NA	NA	grupo 2	NA	NA	NA	NA

Tabla 4.1: Ejemplo de semillas para agrupamiento binario (NA = no asignado).

A manera de ejemplo si tuvieramos los objetos mostrados en la figura 3.3, si se crean las semillas en las posiciones tres y seis, tendríamos el cromosoma mostrado en la tabla 4.1. Una vez aplicado el algoritmo de asignación obtendríamos un agrupamiento binario similar al mostrado en la figura 4.1.

4.1.2. Comparativo de resultados

Se procedió a realizar el experimento descrito en el capítulo anterior para comparar el algoritmo propuesto con los anteriores, en la tabla 4.2 se pueden observar los resultados obtenidos, es importante mencionar que en dicha tabla se han omitido los resultados del algoritmo SICM por cuestiones de espacio, además de que en el capítulo anterior se concluyó que es un algoritmo poco efectivo en la busqueda de la máxima razón de la variabilida entre los

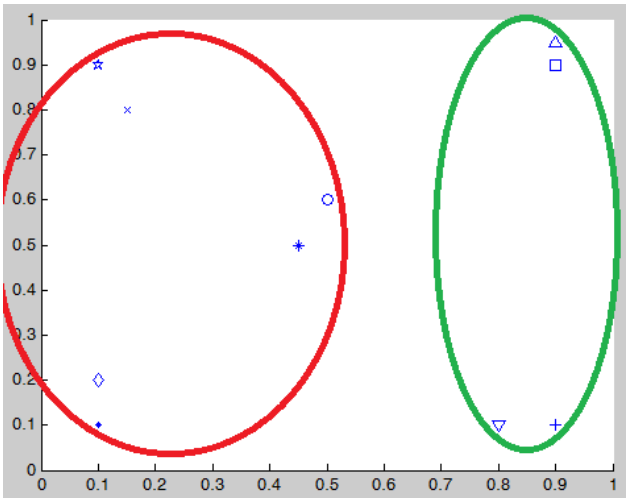


Figura 4.1: Ejemplo agrupamiento binario.

grupos sobre la variabilidad dentro de los grupos.
Se puede observar que el algoritmo propuesto mejora los resultados que obtiene el algoritmo STCM, sin embargo no mejora los resultados obtenidos por AHCM y CSPM. El punto a favor del algoritmo propuesto es que mejora el tiempo de procesamiento a los tres algoritmos genéticos estudiados: SICM, STCM y CSPM. En la tabla 4.3 se pueden observar los tiempos obtenido en la fase de pruebas.

4.1.3. Conclusión

En este capítulo se ha propuesto un algoritmo en base a los métodos previamente analizados, la finalidad de esta propuesta es crear un algoritmo que mejorara los resultados obtenidos por los métodos estudiados. Es por ello que se tomó como base el algoritmo STCM para modificar el funcionamiento del algoritmo genético ya que inicialmente este método tiene un amplio espacio de soluciones donde muchas de estas pueden ser descartadas si se genera una estrategia en la creación de la población, el tener este gran espacio de soluciones ocasiona que el algoritmo encuentre soluciones no tan eficaces además de que tarde mas tiempo en encontrar una solución, es por eso que es deseable ayudarlo a descartar las soluciones erróneas.
Es así como se optó por tomar la idea de la creación de semillas y el algoritmo de asignación que propone el método CSPM, con la finalidad de ayudar al

Number of objects	AHCM F	STCM		CSPM		SACM	
		$\bar{F}_2$ ($\bar{F}_2/F$)	δF_2 (Z_2)	$\bar{F}_3$ ($\bar{F}_3/F$)	δF_3 (Z_3)	$\bar{F}_4$ ($\bar{F}_4/F$)	δF_4 (Z_4)
10	7.14	4.11	0.05	8.79	0.29	4.10	0.11
		(0.57)	(-309.11)	(1.23)	(30.77)	(0.57)	(-147.46)
20	33.74	24.88	2.13	33.22	1.04	27.09	2.83
		(0.73)	(-22.72)	(0.98)	(-2.71)	(0.80)	(-12.83)
30	40.55	33.51	6.15	40.49	2.87	35.97	3.83
		(0.82)	(-6.26)	(0.99)	(-0.11)	(0.88)	(-6.54)
40	31.20	37.50	9.26	55.74	6.64	43.44	7.08
		(1.20)	(3.72)	(1.78)	(20.23)	(1.39)	(9.46)
50	47.08	32.83	4.48	54.04	4.32	39.15	4.35
		(0.69)	(-17.38)	(1.14)	8.81	(0.83)	(-9.97)

Tabla 4.2: Comparativo de efectividad con el método propuesto

Número de objetos	Tiempo (segundos)				
	AHCM	SICM	STCM	CSPM	SACM
10	0.01	0.99	0.67	0.76	0.55
20	0.08	5.73	1.64	4.11	1.44
30	0.34	16.42	3.69	16.18	3.26
40	0.98	24.63	6.15	34.87	5.75
50	2.41	42.59	8.99	64.05	8.00

Tabla 4.3: Tiempo de procesamiento de los métodos estudiados ($N \leq 50$)

algoritmo STCM a descartar las soluciones que proponen un agrupamiento incongruente.
Como resultado de esta propuesta se encontró que el algoritmo propuesto mejora las soluciones del algoritmo STCM, pero no mejora al algoritmo CSPM. Sin embargo el método propuesto mejora el tiempo de procesamiento de los tres algoritmos genéticos estudiados.

4.2. Resultados experimentales

En este capítulo se detalla el procedimiento que se siguió para analizar, los datos de calidad del aire con las técnicas de agrupamiento mencionadas en los capítulos anteriores.

Es importante mencionar que los datos procesados pertenecen al Distrito Federal y Zona Metropolitana del Valle de México (ZMVM). En primera instancia se planteó la idea de analizar los datos correspondientes a la ciudad de Puebla, sin embargo al analizar las bases de datos de calidad del aire se encontraron los siguientes problemas:

- Los datos no cuentan con una organización adecuada para el análisis.
- Se carece de mediciones, debido a fallas en las estaciones de monitoreo ambiental.
- En la ciudad de Puebla solo se cuenta con mediciones de cinco estaciones de monitoreo ambiental.

Otro comentario importante es que en la norma ambiental para el Distrito Federal NADF-009-AIRE-2006, que establece los requisitos para elaborar el Índice Metropolitano de Calidad del Aire (IMECA). La sección "CAMPO DE APLICACIÓN" indica lo siguiente:
"La presente norma aplica en el territorio del Distrito Federal. El IMECA se dará a conocer con base a las zonas de contaminación definidas como Noroeste, Noreste, Centro, Suroeste y Sureste. Su empleo puede extenderse a los municipios conurbados del Estado de México que comprende la ZMVM. [9]"
Es decir, que el cálculo del IMECA solo se aplica a dicha área geográfica. Al momento de finalizar esta tesis no se ha encontrado una norma ambiental que especifique como calcular el IMECA para la ciudad de Puebla. Sin embargo, en caso de existir una especificación de como realizar estos cálculos, las modificaciones a la aplicación desarrollada son mínimas.

4.2.1. Datos de calidad del aire

Los datos a procesar son mediciones tomadas por estaciones de monitoreo ambiental, ubicadas en distintos puntos del D.F. y ZMVM; como la norma NADF-009-AIRE-2006 lo indica, el valor IMECA se reporta todos lo días a cada hora en el sitio web: *http://www.calidadaire.df.gob.mx*, estas mediciones son almacenadas para tener un registro del comportamiento de la calidad del aire en la ciudad de México. En la figura 4.2 se puede observar la ubicación de la estaciones de monitoreo ambiental en la zona del DF y ZMVM, se puede observar que las estaciones de monitoreo se encuentran en funcionamiento, así como la calidad del aire al momento de la consulta y las recomendaciones

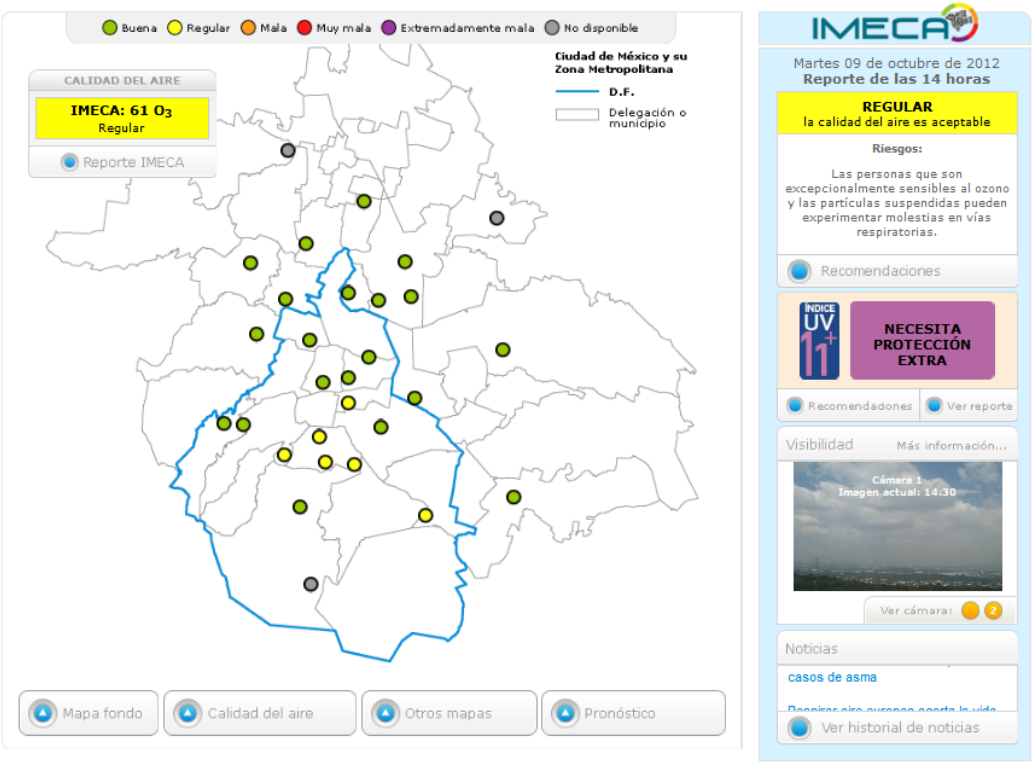
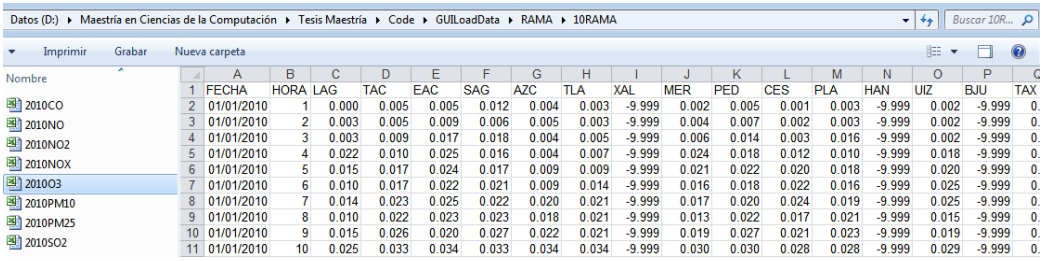


Figura 4.2: Sitio web de monitoreo de calidad del aire del D.F. y ZMVM.

a seguir de acuerdo al estatus del monitoreo.
Hasta el momento de la realización de esta tesis se cuenta precisamente con herramientas de este tipo, donde su principal funcionalidad es tomar las mediciones de cada una de las estaciones de monitoreo ambiental, hacer la correspondiente conversión al valor IMECA de cada uno de los contaminantes criterio y finalmente, mostrar dicho resultado. Además de que se cuentan con recomendaciones de acuerdo al nivel de contaminación del aire, así como un pronóstico de la calidad del aire para el día consultado. Esta información es importante y cumple con lo especificado por la norma NADF-009-AIRE-2006, que dice que se debe notificar a la población acerca del estado de la calidad del aire, así como dar las medidas que se deben de tomar para cada situación dada.
El interés por realizar la presente investigación es para contestar las siguientes preguntas: ¿Existe un patrón en los registros de cada año?, ¿Solo un

contaminante criterio se dispara por medición?, ¿Cuáles son los contaminantes que se disparan con mayor frecuencia?. Estas preguntas no pueden ser contestadas con solo tener el registro de la calidad del aire en un momento determinado, si no que se requiere del análisis de las mediciones de calidad del aire para observar cual es el comportamiento de los datos. Y obtener las conclusiones pertinentes.
Es sumamente importante contar con una organización en los registros de las mediciones a procesar, pues esta organización nos permite manipular los datos en la manera que creamos mas conveniente. Esta es la razón por la que se optó por procesar los datos correspondientes al D.F. Y ZMVM. La aplicación que se mostrará en la siguiente sección funciona leyendo archivos en Excel donde el directorio principal contiene una carpeta para cada año, cada una de estas carpetas se ha nombrado con los dos últimos dígitos del año a procesar y la palabra RAMA, es decir para consultar los archivos del año 2010, es necesario entrar a la carpeta 10RAMA. Dentro de cada año se encuentran los archivos en Excel que son nombrados por el año y el nombre del contaminante criterio, por ejemplo para abrir el archivo de Excel del contaminante ozono, es necesario abrir el archivo en la siguiente ruta 10RAMA/2010O3. Cada archivo Excel a su vez se caracteriza por que en la primera columna se encuentra la fecha de la medición, en la segunda columna la hora y las demás columnas contienen los registros de las estaciones de monitoreo ambiental para el contaminante y año deseados. En la figura 4.3 se muestra un ejemplo de como se encuentran constituidos estos archivos en Excel, así como la ubicación descrita. Es evidente que esta organización nos permite automatizar la lectura de cada uno de los archivos y ser flexibles al momento de mostrar los datos disponibles al usuario.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	FECHA	HORA	LAG	TAC	EAC	SAG	AZC	TLA	XAL	MER	PED	CES	PLA	HAN	UIZ	BJU	TAX
2	01/01/2010	1	0.000	0.005	0.005	0.012	0.004	0.003	-9.999	0.002	0.005	0.001	0.003	-9.999	0.002	-9.999	0.
3	01/01/2010	2	0.003	0.005	0.009	0.006	0.005	0.003	-9.999	0.004	0.007	0.002	0.003	-9.999	0.002	-9.999	0.
4	01/01/2010	3	0.003	0.009	0.017	0.018	0.004	0.005	-9.999	0.006	0.014	0.003	0.016	-9.999	0.002	-9.999	0.
5	01/01/2010	4	0.022	0.010	0.025	0.016	0.004	0.007	-9.999	0.024	0.018	0.012	0.010	-9.999	0.018	-9.999	0.
6	01/01/2010	5	0.015	0.017	0.024	0.017	0.009	0.009	-9.999	0.021	0.022	0.020	0.018	-9.999	0.020	-9.999	0.
7	01/01/2010	6	0.010	0.017	0.022	0.021	0.009	0.014	-9.999	0.016	0.018	0.022	0.016	-9.999	0.025	-9.999	0.
8	01/01/2010	7	0.014	0.023	0.025	0.022	0.020	0.021	-9.999	0.017	0.020	0.024	0.019	-9.999	0.025	-9.999	0.
9	01/01/2010	8	0.010	0.022	0.023	0.023	0.018	0.021	-9.999	0.013	0.022	0.017	0.021	-9.999	0.015	-9.999	0.
10	01/01/2010	9	0.015	0.026	0.020	0.027	0.022	0.021	-9.999	0.019	0.027	0.021	0.023	-9.999	0.019	-9.999	0.
11	01/01/2010	10	0.025	0.033	0.034	0.033	0.034	0.034	-9.999	0.030	0.030	0.028	0.028	-9.999	0.029	-9.999	0.

Figura 4.3: Ejemplo de organización de las mediciones de calidad del aire.

4.2.2. Análisis del dominio

El primer reto para hacer agrupamiento en bases de datos de calidad del aire es la gran cantidad de estos, por cada día se tienen 24 registros de cada contaminante criterio (uno por cada hora), esta información es por cada estación de monitoreo disponible. Por ejemplo, en un año con 365 días tenemos 8760 registros por cada contaminante criterio, en los años como mínimo se tienen registros de 6 contaminantes criterio, estos nos da un total de 56520 mediciones. Además de que tenemos que multiplicar dicho número por el número total de estaciones de monitoreo deseadas.
Antes de aplicar técnicas de agrupamiento se tuvo que construir una aplicación para cargar las mediciones con la finalidad de facilitar la carga de los datos y filtrar la información, esta aplicación permite a los tomadores de decisiones cargar los datos de un periodo de tiempo específico (en total se tienen registros desde 1995 hasta 2010), después de la selección de un periodo de tiempo se deben seleccionar los contaminantes deseados, por ejemplo de 1995 a 2002 se tienen registros para CO, NO_2, NOX, O_3, PM_{10} y SO_2; de 2003 a 2010 existen también registros de los contaminantes NO y PM_{25}. Una vez seleccionados estos parámetros la aplicación habilita las estaciones de monitoreo con datos disponibles para el periodo de tiempo y los contaminantes criterio seleccionados. Estos datos pueden ser graficados para obtener una idea de como se comportan las mediciones. Además de que se les permite a los tomadores de decisiones seleccionar si ellos van a aplicar el agrupamiento con medidas normales, convertirlas a IMECA o a una escala de 0 a 1. La figura 4.4 muestra la aplicación descrita previamente.
Una vez que el tomador de decisiones ha seleccionado los datos, el puede seleccionar un algoritmo de agrupamiento con la finalidad de identificar patrones en los datos.

4.2.3. Desarollo del modelo

Nuestra propuesta es desarrollar un modelo donde podamos identificar el comportamiento de los datos cuando existe una contingencia, la norma mexicana indica que si un contaminante criterio excede la categoría de mala, entonces se declara una contingencia. Sin embargo no existe alguna regla si uno o más contaminantes se encuentran en el intervalo de malo. En nuestra opinión creemos que si existe más de un contaminante en el intervalo de malo también se puede causar daño a los seres vivos y al medio ambiente.

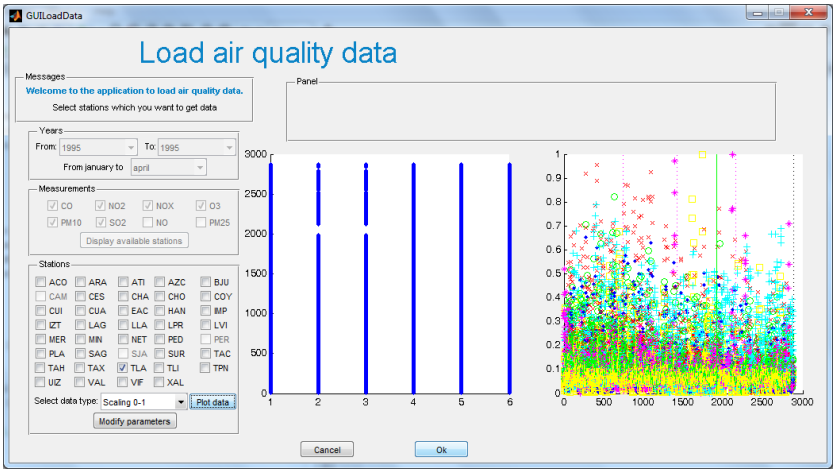


Figura 4.4: Aplicación para la carga de datos de calidad del aire.

Es así como se decidió aplicar técnicas de agrupamiento para identificar como es que los contaminantes criterio se agrupan, además de que podemos determinar cuales son los contaminantes con mayor presencia en la zona metropolitana del valle de México. Esta información es útil en decisiones ambientales, ya que los tomadores de decisiones tienen información de apoyo para tomar las medidas correctas y al mismo tiempo verificar si las acciones que tomaron fueron adecuadas cuando se verifica el comportamiento de los datos una vez que las acciones han sido implementadas.

4.2.4. Resultados

Una vez que la aplicación ha cargado y se han filtrado los datos, podemos aplicar los algoritmos AHCM y CSPM al procesamiento de los datos de calidad del aire, ambas técnicas arrojaron resultados interesantes.
En la figura 4.5, se puede observar una prueba del agrupamiento donde se han agrupado los datos para identificar el porcentaje de contaminantes criterio en los meses de enero a febrero de 1995. El agrupamiento óptimo nos arrojó 10 grupos, es importante mencionar que en esta prueba se escalaron las mediciones a escala de 0 a 1 (eje vertical) con el fin de tener una correspondencia en los datos. En el área obscura se encontró mayor presencia de los contaminantes criterio CO y SO_2 19.99 % y 19.83 % respectivamente. Ambos contaminantes se presentan en un área con baja contaminación. De

estos contaminantes aparecen con mayor presencia en las zonas de baja contaminación (entre el área amarilla y negra). Lo opuesto pasa para NOX y O_3, los cuales tienen mayor presencia en áreas con alta contaminación. De hecho en áreas de magenta a amarilla tienen 44 % de presencia aproximadamente y en el área roja NOX tiene 75 % y O_3 25 %.

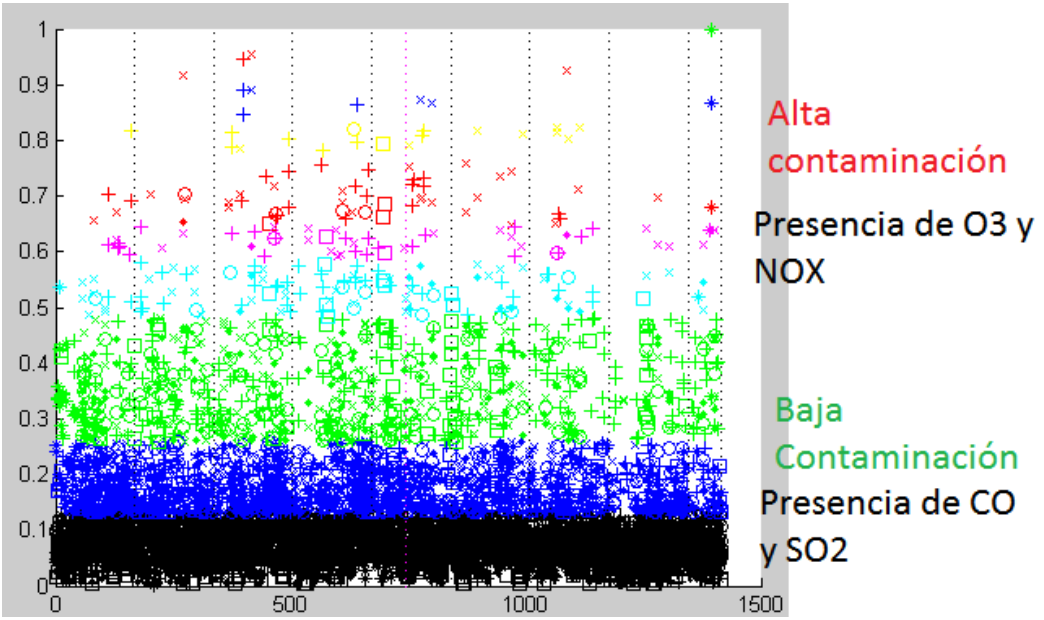


Figura 4.5: Agrupamiento de similaridades en contaminantes criterio.

Otra prueba fue realizada con el objetivo de identificar similaridades en las mediciones de los contaminantes criterio, es decir se tomo como un elemento una medición con todos los contaminantes involucrados y se realizó en agrupamiento basándose en esta comparación.
La figura 4.6 muestra el agrupamiento donde podemos observar tres agrupamientos principalmente, el grupo de color azul contiene las mediciones altas en PM_{10} con media 164.70 IMECA, este agrupamiento es intermitente en el conjunto de datos. Por su parte el grupo en color verde es continuo en el conjunto de datos, sin embargo el mayor valor es O_3 con media 129.13 IMECA. El área de color obscuro es menor que las áreas en color verde y azul, sin embargo esta área contiene el doble de elementos que las otras y es el tipo mas común de medición, el PM_{10} es el valor mas alto con media 67.06 IMECA.

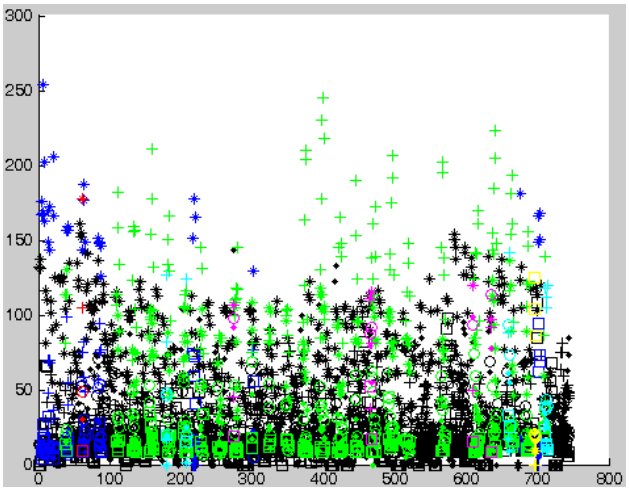


Figura 4.6: Agrupamiento de mediciones contaminantes criterio.

Capítulo 5

Conclusiones y Trabajo a Futuro

5.1. Conclusiones

En la presente tesis se habló acerca de los problemas del cambio climático, haciendo énfasis en el problema de calidad de aire, es así como se presentaron las areas de descubrir conocimiento y cómputo suave. De este modo se analizaron las técnicas de agrupamiento, así como los algoritmos genéticos, ideas que fueron utilizadas para observar que tipo de resultados se pueden obtener con el uso de las técnicas de agrupamiento.
Tomar buenas decisiones en lo que respecta a la contaminación del aire ayuda al medio ambiente y a la salud de los seres humanos. Esta es la principal razón de aplicar técnicas para el análisis de los datos con la finalidad de dar a los tomadores de decisiones información útil que contribuya a los intentos por reducir este problema.
Las técnicas de agrupamiento pueden ayudar a identificar el comportamiento de las mediciones; por ejemplo ellas pueden darnos una idea acerca de los contaminantes comunes en una ciudad. Además de que podemos comprobar si las decisiones tomadas reflejan un resultado en los datos. El estudio de las técnicas de agrupamiento nos llevó a la necesidad de evaluar que tan efectivos son los algoritmos de agrupamiento y hacer una comparación de estos para determinar cual es el algoritmo que nos provee mejores soluciones.
El principal problema al análizar los datos de calidad del aire es que se cuenta con mucha información para procesar, esto ocasiona que las técnicas de

agrupamiento tarden demasiado tiempo en procesar los datos.
Es importante entender los datos y ayudar al algoritmo a encontrar una buena solución, una buena idea es preguntar al tomador de decisiones acerca de su objetivo o sus experiencias. Además, el tener una aplicación para manipular los datos puede ayudarnos a tener una idea acerca de el comportamiento de las mediciones.

5.2. Trabajo futuro

El trabajo a futuro es aplicar otras técnicas para el análisis de los datos con la finalidad de construir un sólido sistema para el soporte a la toma de decisiones, de este modo facilitar el uso de los datos con modelos y estructuras que ayuden a entender el problema de la contaminación del aire.
Usar el algoritmo de análisis de componentes principales(PCA, por sus siglas en inglés) con el fin de reducir las dimensiones en los datos y analizar si esto facilita el procesamiento de los mismos.
Extender la aplicación a otras regiones de México para que la herramienta construida ténga un beneficio a la población.
Programar los algoritmos de agrupamiento para que trabajen en forma paralela con la finalidad de reducir el cómputo de los datos.
Extender la aplicación creada para que pueda ser utilizada desde una aplicación Web.

Bibliografía

[1] The science of climate change: Questions and answers, August 2010.

[2] Anderson, D. R., Sweeney, D. J., and Williams, T. A. *Estadstica para Administración y Economía*. Thomson, 2005.

[3] Bezdek, J. C. *Pattern recognition with fuzzy objective function algorithms.* Plenum Press, 1981.

[4] Bezdek, J. C. *Fuzzy Models for Pattern Recognition: Methods That Search for Structures in Data*. IEEE, 1992.

[5] Chiou, Y.-C., and Lan, L. W. Genetic clustering algorithms. *European Journal of Operational Research 135* (2001), 413–427.

[6] CMNUCC. Convención marco de las naciones unidas sobre el cambio climático.

[7] Department of Sustainability, Enviroment, W. P., and Comunities. Air pollutants.

[8] EPA. Air pollutants.

[9] Federal, D. Gaceta oficial del distrito federal, Noviembre 2006.

[10] Fisher, R. The use of multiple measurements in taxonomic problems. *Annals of Eugenics* (1936), 179–188.

[11] Goldberg, D. E. *Genetic Algorithms in Search, Optimization, and Machine Learning*. Addison-Wesley, 1989.

[12] Hamel, L. *Knowledge discovery with support vector machines*. Jonh Wiley & Sons, Inc., 2009.

[13] Holland, J. H. *Adaptation in Natural and Artificial Systems.* University of Michigan Press, 1975.

[14] Institute, B. The world's worst pollution problems: The top ten of the toxic twety. *Report* (2008), 72.

[15] López, S. *Algoritmo de agrupamiento genético borroso basado en el algoritmo de las C-medias borroso.* PhD thesis, 2001.

[16] Osorio, M. A., Torrijos, T., Sánchez, A., and Arroyo, O. Preliminary analysis for an air quality management dss in the metropolitan valley of puebla, mexico. *11th WSEAS international conference on Wavelet analysis and multirate systems: recent researches in computational techniques, non-linear systems and control* (2011), 210–215.

[17] Sánchez, A. *Soft Computing.* BUAP, 2011.

[18] Sierra, B. *Aprendizaje Automático: Aspectos práctivos utilizando el software WEKA.* Pearson Educación, 2006.

[19] Ward, J. H. Hierarchical grouping to optimize an objetive function. *Journal of the American Statiscal Association 58* (1963), 236–244.

[20] Wikipedia. Cambio climático.

[21] Wikipedia. Contaminación atmosférica.

[22] Wikipedia. Iris flower data set.

[23] Wikipedia. Soft computing.

[24] Zadeh, L. A. Fuzzy logic, neural networks, and soft computing. *Communication of the ACM 37* (March 1994), 77–84.