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 cómputo 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 [Wikipedia(b)].

Definición 1. *(Cambio climático) [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.*

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) [Aus(2010)]. El cambio climático es una variación en el patrón promedio del clima sobre un largo periodo de tiempo .*

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 [Aus(2010)].

¿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 dramáticamente, 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 atmósfera, la intensidad del sol y la órbita 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 órbita 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 recientes 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 atmósfera. 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. Después 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 [Wikipedia(a)].
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.

Contaminantes aéreos

Los contaminantes aéreos pueden ser clasificados de manera general en dos categorías [Institute(2008)]:

Contaminantes primarios son aquellos que son emitidos a la atmósfera 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 atmósfera cuando los contaminantes primarios reaccionan con la luz del sol, oxigeno, agua y otros químicos presentes en el aire.

Además de esta clasificación los contaminantes aéreos pueden ser encontrados en ambientes exteriores e interiores, estos pueden ser divididos en tres grupos [Department of Sustainability and Comunities.()]:

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

Esta clasificación es más especializada que la anterior, a continuación se describen cada una de las categorías antes mencionadas, es importante resaltar que en esta tesis trabajamos con los contaminantes criterio, a esto se debe que se describan a mayor detalle.

Contaminantes criterio

Contaminantes criterio es un término 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 [Department of Sustainability and Comunities.()]. 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 [EPA()].

Ozono (O_3): No es emitido directamente en el aire, pero es creado por reacciones químicas entre óxidos 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 "Óxidos 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 "Óxidos 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 [Osorio et al.(2011)Osorio, Torrijos, Sánchez, and Arroyo]. 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 qué información relevante podemos encontrar en estos.

Definición 3. *(Descubrir conocimiento [Knowledge Discovery]) [Hamel(2009)]. Descubrir conocimiento es un proceso semi automático para extraer información útil de colecciones de datos que son demasiado grandes como para ser investigados manualmente.*

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 término usualmente asociado con descubrir conocimiento es *minería de datos (data mining)*. La minería de datos puede ser considerada como una forma de descubrir conocimiento, ésta tiene como objetivo extraer información de bases de datos; la minería de datos es usualmente referida cómo descubrir conocimiento en base de datos *(knowledge discovery in databases (KDD))* [Hamel(2009)].
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.

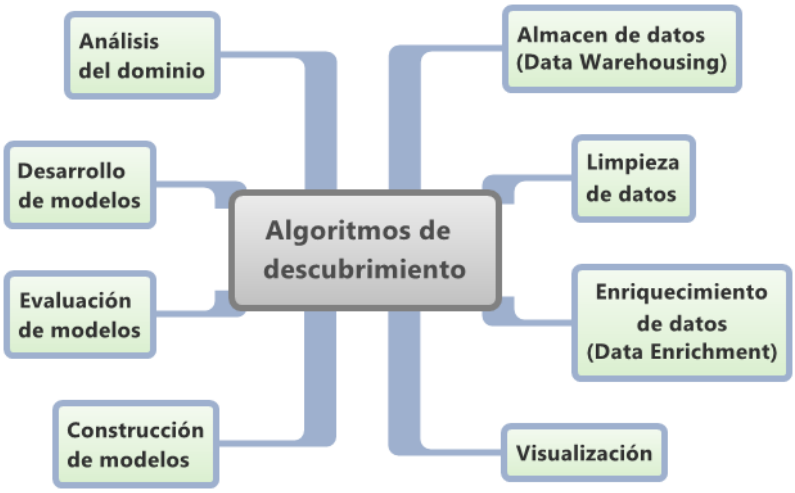

Figura 1.1: Procesos para descubrir el conocimiento.

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 alguna 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 presentan 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 una corazonada" acerca de los datos e identificar la su calidad. 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 centramos 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 artificial 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 técnicas 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 significa simplemente predecir el valor del atributo objetivo para un conjunto de objetos. En otros casos 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 [Hamel(2009)].

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 infor-

mación que se tiene almacenada.
Según algunos autores, el objetivo básico de todas las ciencias es el reconocimiento de patrones [Bezdek(1981)].
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 [Bezdek(1992)]. 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 [López(2001)]:

- 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 [López(2001)]:

   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 automóvil 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 [Sánchez(2011)].
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]) [Zadeh(1994)]. 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 [Zadeh(1994)], las técnicas que conforman esta área son [Wikipedia(d)]:

- Redes neuronales
- Sistemas difusos

[1] Lo que se denomina "Aprendizaje supervisado"

- 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):

Definición 5. *(Cómputo suave) [Zadeh(1994)]. 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 cómputo 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.