**CLASIFICACIÓN DE DATOS**

Con la evolución de la computación, las redes sociales, la búsqueda e información en la red y el avance tecnológico en cada una de las diferentes áreas de conocimiento, cada día surgen billones de bytes de datos nuevos. Por ejemplo, la red social Facebook aporta diariamente 600 terabytes (1 terabytes= un billón de bytes), los sistemas de almacenamiento satelital crea a cada instante millones y millones de datos de imágenes e información meteorológica en tiempo real. Si a esto le sumamos que estos datos pueden cambiar rápidamente en el tiempo y ser estructurados y no estructurados, además de estar presentados en diferentes formas: datos multimedia, datos de red, datos de series temporales, archivos de texto, vídeos, audios, datos en bases relacionales, registros de red, gráficos geoespaciales en 3D, mapas, datos médicos, datos biológicos, entre otros. Tener toda esta información genera retos importantes en el sentido de obtener conocimiento a partir de la misma y dar sentido a la información que se tiene. Una de las herramientas de la minería de datos y la estadística en el manejo de la información es la clasificación de los datos y ha sido uno de los campos más estudiados en los últimos años.

Por esta razón se hace necesario encontrar mecanismos que permitan clasificar un cierto grupo de datos en categorías o subgrupos donde los datos de cada de una de estas este determinado por una propiedad. Asimismo se busca crear algoritmos que permitan determinar a qué categoría pertenece un nuevo dato u observación, partiendo de que a priori se conocen las subcategorías de la base de datos, es decir, conocemos las etiquetas que asignamos en cada una de las particiones. Esto es lo que se conoce con el nombre de ***Aprendizaje Supervisado.*** Por ejemplo, si consideramos los correos electrónicos, podemos determinar dos categorías a las cuales le podemos asignar “spam” y “no spam” y de esta manera determinar mediante un algoritmo a cuál categoría pertenece un nuevo correo electrónico o email. En el campo de la medicina, por ejemplo, podríamos a un paciente a partir de una serie de características determinadas como: sexo, edad, presión, presencia de síntomas etc. En resumen podemos decir que la clasificación en una herramienta que permite reconocer patrones.

Por medio de la clasificación de los datos un meterológo, por ejemplo, podría realizar con mayor exactitud predicciones sobre el clima, indicando si será lluvioso, soleado o nublado.

Por ejemplo, el algoritmo Bayesiano, en el campo de la clasificación no supervisada, es usada para categorizar los correos electrónicos que comúnmente denominamos como spam. Día a día la clasificación esta presente en cada una de las actividades que realizamos

La clasificación es un tema muy amplio y cada uno de los algoritmos diseñados para tal fin depende en gran medida de la característica de los datos, del área del conocimiento que estemos tratando y del problema que se quiere resolver.

Dentro de los métodos básicos para clasificación de información tenemos:

* Clasificación probabilística
* Arboles de decisión
* Métodos basados en reglas
* Métodos SVM
* Redes neuronales

Otras variaciones de los métodos anteriores que mejoran la calidad de los resultados son:

* Método de conjunto
* Métodos visuales
* Aprendizaje de transferencia
* Métodos semi-supervisados
* Aprendizaje activo

**Clasificación supervisada:** Este tipo de clasificación se caracteriza por que se tiene un conocimiento a priori de las categorías o clases en las que vamos a dividir nuestro conjunto de datos, es decir, los objetos están agrupados por características comunes. Para hacer esta clasificación se usa una función denominada ***clasificador*** que asigna a los datos de entrada una respectiva categoría y de esta manera permite segmentar el gran grupo de datos en subgrupos con patrones en común, adicionalmente, esto permite determinar a cual subcategoría un elemento que no esta dentro del conjunto inicial, es decir, permite clasificar nuevos datos. Este tipo de clasificación la podemos usar en diagnostico de enfermedades, predicciones financieras, reconocimiento de caracteres escritos a mano, entre otros. Dentro de este grupo podemos encontrar los siguientes algoritmos

**Clasificación no supervisada:** En este caso no se cuenta con el conocimiento a priori de las subcategorías o etiquetas en los diferentes grupos en los que se pretenden agrupar los datos. A este tipo de clasificación es a lo que se le conoce usualmente como *“clustering”*  y su objetivo fundamental es el descubrimiento de grupos de “objetos” cuyas características afines permitan separarlos en diferentes clases.

Si no existe el conocimiento acerca de las etiquetas de los patrones, entonces para clasificar un objeto es necesario de un proceso previo de análisis de los datos que se conoce como clasificación no supervisada o técnicas de agrupamiento (clustering) los cuales proporcionan información sobre la estructura de los datos. El objetivo de hacer un agrupamiento es clasificar un conjunto de objetos en subconjuntos o grupos en los cuales los elementos de cada clase guardan un alto grado de semejanza y que un objeto que no esté en determinada clase difiera bastante de los elementos de la clase.

El aprendizaje no supervisado realza su importancia cuando se dispone de conjuntos de datos en los cuales no hay etiquetas que permitan hacer una subclasificación, o cuando los patrones para determinar cuando un elemento esta o no en determinada clase varian con el tiempo. En estos casos es necesario procesar los datos para luego clasificarlos. La forma de hacer esto es por medio de algoritmos de agrupamiento.

El agrupamiento es una poderosa herramienta que se usa en el minería de textos, procesamientos de secuencias, recuperación de información, segmentación de imágenes, compresión de datos, procesamientos de datos espaciales, clasificación de zonas geográficas, análisis de imágenes de satélites, visualización de datos, organización de documentos en una biblioteca, entre muchos otros contextos donde se manejen grandes volúmenes de datos.

La clasificación es una forma de análisis de datos que extrae modelos que describen clases de datos importantes. Estos modelos, llamados clasificadores, predicen etiquetas categóricas (discretas, no ordenadas). Por ejemplo, podemos construir un modelo de clasificación para clasificar las solicitudes de préstamos bancarios como seguras o riesgosas. Este tipo de análisis puede ayudar a comprender mejor los datos en general. Muchos de los métodos de clasificación han sido propuestos por los investigadores en el aprendizaje automático, el reconocimiento de patrones y las estadísticas. La mayoría de los algoritmos son residentes de la memoria, por lo general suponiendo un pequeño tamaño de datos. La investigación reciente de minería de datos se ha basado en dicho trabajo, desarrollando una clasificación escalable y técnicas de predicción capaces de manejar grandes cantidades de datos residentes en disco. La clasificación tiene numerosas aplicaciones, incluyendo detección de fraude, marketing objetivo, predicción de desempeño, fabricación y diagnóstico médico.

En la terminología del aprendizaje automático, [1] la clasificación se considera una instancia de aprendizaje supervisado, es decir, aprendizaje en el que se dispone de un conjunto de formación de observaciones correctamente identificadas. El procedimiento no supervisado correspondiente se conoce como agrupación, e implica la agrupación de datos en categorías basadas en alguna medida de similitud inherente o distancia.  
  
A menudo, las observaciones individuales se analizan en un conjunto de propiedades cuantificables, conocidas de diversas maneras como variables explicativas o características. Estas propiedades pueden ser diversas (por ejemplo, "A", "B", "AB" u "O" para el tipo sanguíneo), ordinal (por ejemplo, "grande", "medio" o "pequeño") El número de ocurrencias de una palabra en particular en un correo electrónico) o de valor real (por ejemplo, una medición de la presión arterial). Otros clasificadores funcionan comparando observaciones con observaciones anteriores por medio de una función de similitud o distancia.  
  
  
  
La terminología entre campos es muy variada. En las estadísticas, donde la clasificación se realiza a menudo con regresión logística o un procedimiento similar, las propiedades de las observaciones se denominan variables explicativas (o variables independientes, regresores, etc.) y las categorías a predecir se conocen como resultados que se consideran como Valores posibles de la variable dependiente. En el aprendizaje automático, las observaciones se conocen con frecuencia como instancias, las variables explicativas se denominan características (agrupadas en un vector de características) y las posibles categorías a predecir son clases. Otros campos pueden utilizar terminología diferente: p. En la ecología de la comunidad, el término "clasificación" normalmente se refiere al análisis de conglomerados, es decir, a un tipo de aprendizaje no supervisado, en lugar del aprendizaje supervisado descrito en este artículo.