Inteligencia Artificial

Sistema de Clasificación de Contenido Web

Alumno: Carlos Emanuel Balcazar
Email: emanuelbalcazar13@gmail.com
Proyecto: https://github.com/emanuelbalcazar/web-classifier

Profesores:
Dr. Claudio Delrieux.
Lic. Romina Stickar.

Resumen

El aprendizaje automático es un subcampo de las ciencias de la computación y una rama proveniente de la inteligencia artificial en donde se pretende desarrollar técnicas que permitan a las computadoras aprender. La forma en la que se realiza es desarrollando programas que son capaces de inducir el conocimiento a partir de una muestra de información suministrada como ejemplo.

El aprendizaje automático es ampliamente utilizado en una amplia cantidad de aplicaciones entre los cuales están los motores de búsqueda, diagnósticos médicos, detección de intrusiones en la red, clasificación de datos, reconocimiento del habla y el lenguaje escrito, juegos, robótica, etc.

Con el fin de aprovechar la potencia de este tipo de aplicaciones, se desarrollara una aplicación web que permita clasificar páginas web en función de su contenido, además de brindar herramientas que permitan usar y comparar los diferentes algoritmos de aprendizajes existentes para conocer su efectividad y precisión.

Objetivos

El presente documento pretende abarcar la información preliminar y todos los detalles pertinentes al desarrollo de un sistema de clasificación de páginas webs. Dicho documento contendrá toda la información asociada a las temáticas de inteligencia artificial tratadas.

Por otra parte se agregara información técnica asociada al desarrollo realizado, pero no se profundizará en otras cuestiones que tengan que ver directamente con los temas tecnológicos y herramientas.

Motivación

Las temáticas de minería de datos y algoritmos de aprendizaje son en particular, temas muy interesantes en el área de la informática y después de cursar la materia de Inteligencia Artificial y tener acercamientos a dichos temas en el ámbito laboral quede interesado en investigar y realizar una primera aproximación del tipo de desarrollo que me gustaría realizar. La idea general fue propuesta por el profesor a cargo de la materia como posible trabajo y fue de mi agrado poder involucrarme y poder entrar en contacto con un área tan extensa e increíble de la informática.

Marco Teórico

Algoritmos de Aprendizaje

Los sistemas de aprendizaje automáticos intentan imitar la intuición del ser humano para adquirir conocimientos en base los datos existentes y experiencias previas, si bien la intuición humana no puede ser reemplazada en su totalidad, se puede establecer un marco de colaboración entre un experto y la computadora de forma que trabajen en conjunto.

Para ello el aprendizaje automático tiene como resultado un modelo para resolver una determinada tarea entre los cuales se distinguen:

- Los **modelos geométricos** construidos en un espacio que pueden tener múltiples dimensiones, esto modelos se trabajan en un plano mediante cálculos matemáticos aplicados a planos de n-dimensiones.

- Los **modelos probabilísticos** que determinan la distribución de probabilidades que enlace a los valores de los datos con valores predeterminados, uno de los conceptos claves para los modelos probabilísticos es la estadística bayesiana.

- Los **modelos lógicos** que expresan las probabilidades en reglas organizadas en forma de árboles de decisión.

Los diferentes algoritmos de aprendizaje automático se agrupan en función de la forma en la que trabajan, entre ellos están:

- **Aprendizaje supervisado:** el algoritmo establece una correspondencia entre las entradas y las salidas deseadas del sistema. Un tipo de algoritmo es la clasificación, donde el sistema de aprendizaje trata de etiquetar/ clasificar una serie de entradas utilizando como soporte etiquetas previamente indicadas denominada base de conocimiento. Este tipo de aprendizaje es sumamente útil en problemas de investigación biológica y bioinformática.

- **Aprendizaje no supervisado:** todo el proceso de modelado se realiza sobre un conjunto de entradas al sistema de la cual no se tiene información sobre las categorías de esos ejemplos. Por lo tanto el sistema debe ser capaz de reconocer patrones para poder etiquetar las nuevas entradas.

- **Aprendizaje semi supervisado:** es una combinación de los dos algoritmos previos en donde se tienen en cuenta los datos etiquetados.

- **Aprendizaje por refuerzo:** el algoritmo aprende observando el entorno que le rodea en donde su información de entrada es la retroalimentación que obtiene del entorno como respuesta a sus acciones, podemos decir que el sistema aprende mediante el ensayo – error.

Minería de Datos

La minería de datos es un campo de la estadística y de las ciencias de la computación referido al proceso de descubrir patrones en grandes volúmenes de conjuntos de datos. Para ello se suelen utilizar métodos de la inteligencia artificial, aprendizaje automático, estadística y sistemas de bases de datos.

El objetivo general del proceso de minería de datos consiste en extraer la información de un conjunto de datos y transformarla en una estructura comprensible para su uso posterior.

El proceso típico de minería de datos consiste en los siguientes pasos:

- **Selección del conjunto de datos:** se selecciona las variables objetivo (aquellas que se quiere predecir, calcular o inferir) como las variables independientes (usadas para realizar el cálculo o proceso) así como los posibles registros o muestras de datos.

- **Análisis de las propiedades de los datos:** generalmente se busca la presencia de datos atípicos o nulos.

- **Transformación del conjunto de datos de entrada:** se realiza de diversas formas con el objetivo de prepararlo para aplicar la técnica de minería de datos que más convenga en función del problema.

- **Aplicar minería de datos:** se construye el modelo predictivo, de clasificación o segmentación.

- **Extracción del conocimiento:** mediante alguna técnica se obtiene el conocimiento que representa los patrones de comportamiento observados en los valores del problema o las relaciones asociadas entre dichos valores. Esto quiere decir que se logra inferir el conocimiento a partir del resultado de procesar los datos extraídos.

- **Interpretación y validación:** una vez obtenido el modelo resultante se debe validar verificando si las conclusiones obtenidas son válidas y satisfactorias. En el caso de haber obtenido varios modelos mediante el uso de distintas técnicas se deben comparar en busca de aquel que mejor se ajuste al problema.

Las técnicas de minería de datos utilizadas provienen de la inteligencia artificial y la estadística como se mencionó anteriormente, estas técnicas no son más que algoritmos que se aplican sobre un conjunto de datos para obtener resultados.

Las técnicas más representativas son:

- **Redes neuronales:** son un paradigma de aprendizaje y procesamiento automático de datos en donde existe una interconexión de neuronas en una red que colabora para producir una salida en función de un denominado estímulo.

- **Regresión lineal:** es la más utilizada para formar relaciones entre los datos, sin embargo no es suficiente en espacios donde puedan relacionarse más de dos variables en múltiples dimensiones.

- **Arboles de decisión:** es un modelo utilizado en el análisis predictivo en donde se representan una serie de condiciones que suceden de forma sucesiva para la resolución de un problema.

- **Modelos estadísticos:** se expresan en forma de igualdad o ecuaciones en donde se indican que factores alteran la variable de respuesta.

Clasificación de Documentos

Se define la clasificación de documentos como la actividad de etiquetar documentos en lenguaje natural con categorías de un conjunto predefinido. El enfoque actual adoptado para la categorización de documentos se basa en técnicas de aprendizaje automático, se construye un clasificador mediante un aprendizaje inductivo a partir de documentos previamente clasificados.

Los algoritmos de clasificación existentes (native bayes, c4.5, aprendizaje de vectores, etc.) pueden entrenarse para clasificar documentos dados un conjunto grande de ejemplos de entrenamiento en donde cada uno ha sido etiquetado anteriormente en la categoría correspondiente.

Entre los algoritmos de clasificación más conocidos se utilizaron en particular dos para este proyecto:

- **Clasificador bayesiano:** es un clasificador probabilístico fundamentado en el teorema de Bayes en donde se asume que la presencia o ausencia de una característica particular no está relacionada con la presencia de cualquier otra característica. Es decir la presencia de ciertas características aporta a la probabilidad de que un determinado elemento sea clasificado como tal, esta clasificación se puede entrenar de manera eficiente en un entorno de aprendizaje supervisado. La ventaja de este algoritmo es que requiere de una pequeña cantidad de datos de entrenamiento para estimar los parámetros necesarios para la clasificación.

- **Clasificador por Regresión Logística:** este clasificador se apoya en un análisis matemático utilizado para predecir el resultado de una variable que puede adoptar un número determinado de categorías en función de las variables independientes (también llamadas predictoras). Este algoritmo tiene similitud con el anterior mencionado debido a que también hace uso de la probabilidad de un evento.

Es importante destacar que ambos algoritmos son de **aprendizaje supervisado**, es decir se le proveen una serie de documentos ya clasificados por lo cual deben usar de referencia para poder realizar la clasificación correspondiente. La elección de estos dos algoritmos se debió a su implementación ya existente, además de que se deseaba comparar dos algoritmos similares para conocer su eficiencia.

Desarrollo Realizado

Para el desarrollo de la aplicación se tomaron las siguientes decisiones:

- El sistema deberá poder clasificar contenido web, en lo posible analizando los textos que componen algún recurso.
- Se deberán usar dos o más implementaciones a elección de algoritmos de clasificación.
- Los resultados mostrados deben tener el porcentaje de clasificación de cada clasificador para conocer y comparar sus diferencias.
- Se debe tener una base de conocimiento previo en donde se puedan entrenar los algoritmos para ser utilizados.

Herramientas Utilizadas

Las herramientas que se utilizaron para el desarrollo del programa fueron:

- NodeJS – framework JavaScript para el desarrollo de aplicaciones en el lado del servidor.
- Natural - librería de nodejs que posee funciones de clasificación, análisis de sentimientos, cálculo de distancia entre palabras, tokenización, etc.

El proyecto se puede descargar de GitHub junto a sus instrucciónes de instalación.