Predicción de Precipitación

Ricardo Apú Chinchilla[1], Oscar Corella Quirós[2] y Alonso Mondal Durán[3]
CI-1441 Paradigmas computacionales
Escuela de Ciencias de la Computación e Informática
Facultad de Ingeniería
Universidad de Costa Rica
[1]*ricardo.apu@ucr.ac.cr*, [2]*oscar.corella@ucr.ac.cr*, [3]*alonso.mondal@ucr.ac.cr*

Noviembre del 2017

Resumen

En este trabajo se utilizó el algoritmo C5.0 para generar un árbol de decisión con base en datos recolectados de diferentes estaciones metereológicas en Costa Rica. Este árbol es capaa de predecir si va a llover en la siguente hora dados los datos de una hora particular con 89 % de precición.

Palabras clave: árbol de decisión, C5, predicción, clima, algoritmos, clasificación de datos

1. Introducción

Predecir la precipitación ha sido de gran interés para los seres humanos a través de los años. Desde mejorar las cosechas, prevenir inundaciones, a salir con ropa adecuada de la casa, entre muchos otros casos son algunas de las utilidades de esto. Este proyecto consta de un análisis de datos climáticos de diferentes estaciones ubicadas en Costa Rica, para poder determinar y predecir si en algún dado momento lloverá o no. Para esto, se decidió utilizar el algoritmo C5.0 para generar varios árboles de decisión con diferentes combinaciones de los datos provistos para decidir cuál sería más efectiva. Este proceso se separó en varias partes ya que los datos no sólo necesitaban ser analizados, sino previamente curados. En este documento se explican las diferentes etapas del proceso además de demás detalles necesarios para su realización. Este documento incluye la descripción del problema, un breve marco teórico, los objetivos y cronograma del equipo de desarrollo a completar, la propuesta de solución del problema, su desarrollo y experimentación y por último los resultados y análisis del proceso.

2. Problema

Dado un conjunto de atributos metereológicos se quiere poder predecir si va a llover o no con nuevos datos. Existen diferentes algoritmos de clasificación que funcionan para atacar este problema y dar una respuesta binaria al él. Estos algoritmos utilizan datos para crear un árbol de decisión y dar las respuestas. En este proyecto se cuenta con un conjunto de datos de diferentes estaciones metereológicas del país y es necesario curarlos y organizarlos de mejor manera para tratar de optimizar el árbol de decisión. También, los datos vienen en intervalos de treinta minutos y traen muchos campos nulos, lo que complica el procesamiento de todos para generar el árbol.

3. Marco Teórico

3.1. Algoritmo C5.0

El Algoritmo C5 o C5.0 pertenece a una sucesión de constructores de árboles de decisión que remon-

tan sus orígenes al trabajo de Hunt y otros a fines de la década de 1950 y principios de la década de 1960 (Hunt 1962). Sus predecesores inmediatos fueron ID3 (Quinlan 1979)- un sistema simple que consistía inicialmente de unas 600 líneas de Pascal-, C4 (Quinlan 1987) y C4.5 (Quinlan 1993) commo dice el artículo de Kohavi and Quinlan (1999). Las mejoras de C5 sobre sus predecesores abarcan eficiencia y precición del árbol volviéndolo la perfecta opción para generar un árbol de decición ahora (Pandya and Pandya (2015)).

3.2. Predicción Climática

En esta sección se explicarán los datos de más relevancia utilizados para crear la clasificación y análisis de datos.

- Temperatura (máxima, mínima, promedio)
- Humedad relativa
- Velocidad del viento (escalar, media)
- Presión Barométrica
- Radiación solar (máxima, mínima, promedio)

resumen de la revisión bibliográfica y del estudio de propuestas para la resolución de problemas similares; separado en subsecciones, una para cada estudio o enfoque estudiado.

4. Objetivos y cronograma

4.0.1. Objetivo general

Definir un árbol de decisión que permita determinar si va a llover o no dado ciertos atributos climatológicos

4.0.2. Objetivos específicos

- Curar los datos y tomar atributos de interés.
- Analizar el árbol de decisión generado por C5.0 desde el punto de vista del porcentaje de aciertos.

5. Propuesta de solución

Se propone agarrar y clasificar los datos que se tienen de acuerdo a criterios arbitrarios de cuáles son útiles y cuales no. Una vez hecho esto, se separan los datos en datos para entrenar y datos para probar.

6. Desarrollo, prueba y validacion

Primeramente se van a curar los datos ya que hay algunas entradas que vienen con mucha información vacía y otras entradas que no decían si llovió o no lo cual no sirve para los objetivos del proyecto. Esto se debe hacer de manera manual. Una vez que los datos hayan sido correctamente curados, se toman los atributos que parezcan más relevantes y se promedia la información en intervalos de una hora. Seguidamente, se clasifican los datos viendo si llovió o no en la siguiente hora. Para ello, se realizó un script en python.

Para generar el árbol de decision se utiliza el algoritmo de C5.0 creado por el autor Ross Quinlan. Hay que compilar el código de C5.0, y luego ejecutar ese comando con la opción -f para pasarle 3 archivos como parámetro. Uno de los archivos contiene datos para generar el árbol de decisión, otro con datos para probarlo después de que haya sido construido y un tercer archivo que describe los tipos de los datos que se están utilizando.

7. Experimentación y análisis

descripción de experimentos y los resultados obtenidos con la aplicación de la solución desarrollada.

8. Problemas abiertos y problemas futuros

9. Agradecimientos

9.0.1. El paquete natbib

9.0.2. Citas en el texto

Cada vez que deba citar en el texto una o más de las fuentes bibliográficas revisadas, use el comando **citep**. Por ejemplo, la cita (**?**) se escribe así:

```
\citep{minsky1975}
```

Esto hace referencia a la entrada **minsky1975** en el archivo separado que contiene la bibliografía: bibliografía.bib.

9.0.3. El archivo bibliografía.txt

Este archivo debe contener una entrada en el lenguaje **BibTeX** para cada cita bibliográfica diferente en su artículo. En el Apéndice A se muestra el contenido completo de la bibliografía de este artículo.

Puede ver las reglas generales de **BibTeX** para producir cada entrada de la bibliografía en un documento **LaTeX** en (**?**).

9.0.4. La sección Referencias

LaTeX genera automáticamente la lista de referencias bibliográficas en una sección separada al final del artículo. Vea la sección **Referencias** de este artículo para ver cómo aparecen la cita del ejemplo anterior y estos otros ejemplos: (**?**), (**?**).

Referencias

Ron Kohavi and Ross Quinlan. Decision tree discovery, 1999. URL `http://ai.stanford.edu/ ronnyk/treesHB.pdf`.

Rutvija Pandya and Jayati Pandya. C5.0 algorithm to improved decision tree with feature selection and reduced error pruning. *International Journal of Computer Applications*, 117(16):0975–8887, 2015. doi: http://research.ijcaonline.org/volume117/number16/pxc390331

Apéndice A El archivo "bibliografia.txt"

```
@article{fenn2006,
    author  =   {Fenn, J},
    title   =   {Managing Citations and Your Bibliography with BibTeX},
    journal =   {The PracTeX Journal},
    note     =  {\url{http://www.tug.org/pracjourn/2006-4/fenn/}},
    year    =   {2006}
}

@book{mehlhornsanders2008,
    author  =   {Mehlhorn, K and Sanders, P},
    title   =   {Algorithms and data structures. The basic toolbox},
    publisher   =   {Springer, Berlín, Alemania},
    year    =   {2008},
    pages   =   {33--35}
}

@book{minsky1975,
    author    = {Minsky, M},
    title     = {The Society of Mind},
    publisher = {Simon and Schuster, California, EE.UU.},
    year      = {1975},
    pages     = {12--15}"
}

@book{michalskietal1983,
    author    = {Michalski, R S and Carbonell, J G and Mitchell, T M (Eds.)},
    title     = {Machine learning, an artificial intelligence approach},
    volume    = {I},
    publisher = {Springer, Berlín, Alemania},
    year      = {1983}
}
```