

Universidad de La Habana
Facultad de Matemática y Computación



Mitigación de sesgos con ensembles y optimización multiobjetivo

Autor:

Jorge Mederos Alvarado

Tutores:

Juan Pablo Consuegra Ayala
Alejandro Piad Morfis

Trabajo de Diploma
presentado en opción al título de
Licenciado en (Matemática o Ciencia de la Computación)

Fecha

github.com/jmederosalvarado/thesis

Introducción

En la actualidad los algoritmos de aprendizaje automático están siendo aplicados en disímiles áreas de la vida humana. Es común encontrarlos aplicados en sistemas de recomendación de compras, aplicaciones de citas, solicitudes de préstamos, contratación personal y muchas otras áreas. A raíz de ello, ha surgido un creciente interés en estudiar las potencialidades y limitaciones de los modelos de aprendizaje automático, así como las posibles implicaciones de confiar ciegamente en sus predicciones.

En particular, su incorporación a tareas de toma de decisiones de alto riesgo ha dirigido la atención de muchos investigadores hacia una nueva interrogante: ¿están siendo "justos" los algoritmos de aprendizaje automático al tomar sus decisiones?

En este escenario, ha ganado popularidad el desarrollo de técnicas para detectar y mitigar los sesgos en colecciones de datos y algoritmos de aprendizaje automático. Tales herramientas son cruciales para desarrollar sistemas de toma de decisiones más justos. Los estudios orientados hacia la equidad en algoritmos de aprendizaje automático se enfocan principalmente en desarrollar técnicas que consideren tanto la precisión como la equidad de los modelos.

Motivación

Un modelo de aprendizaje de máquina se entrena con el objetivo de optimizar una única métrica, en la mayoría de los casos la precisión. Esto significa que los modelos aprenden muy bien los patrones que se presentan en los datos de entrenamiento, incluyendo aquellos patrones que representan sesgos y prejuicios que están desafortunadamente presentes en la sociedad y por ende en los datos recopilados, en algunos casos incluso amplifican estos patrones negativos. Son varias las técnicas que se han explorado para resolver este problema, algunas se enfocan en un preprocesamiento de los datos para eliminar aquellos elementos que puedan inducir un sesgo en el modelo, otras realizan variaciones en el método de entrenamiento con el mismo objetivo. Sin embargo permanece relativamente poco explorado el uso de técnicas de optimización multiobjetivo que permitan al modelo optimizar hasta encontrar un buen balance entre cuán justo es y cuán preciso.

Otra tecnica que ha demostrado ser de gran utilidad en la prevencion de los sesgos en los modelos de aprendizaje de maquina es la construccion de ensamblados de multiples modelos que maximizan la varianza entre si, por lo que se minimiza el sesgo del ensamblado final.

Problematica

A pesar de que existe AutoGOAL, una biblioteca de AutoML, que permite obtener modelos para resolver problemas arbitrarios utilizando entre otras tecnicas aprendizaje de maquina. No existe una biblioteca o herramienta que permita resolver de principio a fin un problema de clasificacion utilizando aprendizaje de maquina y donde exista alguna garantia de que el modelo aprendido sea justo.

Objetivo general

Proponer una herramienta que permita resolver problemas de clasificacion utilizando aprendizaje de maquina y que permita garantizar que el modelo aprendido sea justo.

Objetivos especifico

- Encontrar modelos que maximicen la varianza para minimizar el sesgo.
- Metodos basados en metaheurísticas para optimizar los modelos utilizando simultaneamente metricas de equidad y precision.
- Explorar adición de optimización multiobjetivo a AutoGOAL para que el modelo aprendido sea justo.
- Metodos basados en la combinacion de diferentes metricas en una sola, para poder aprovechar los multiples metodos de optimización que existen.