

Universidad de La Habana
Facultad de Matemática y Computación



Mitigación de sesgos con ensembles y optimización multiobjetivo

Autor:

Rodrigo García Gómez, Jorge Mederos Alvarado

Tutores:

**Juan Pablo Consuegra Ayala
Alejandro Píad Morfís**

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

2022-06-30

github.com/pat-and-mat/bfair-report

Resumen

Resumen en español

Abstract

Resumen en inglés

Índice general

Introducción	1
1. Estado del Arte	4
2. Propuesta	5
3. Detalles de Implementación y Experimentos	6
Conclusiones	7
Recomendaciones	8

Índice de figuras

Ejemplos de código

Introducción

El aumento de la importancia y el área de uso de los algoritmos de aprendizaje automático ha alcanzado un nivel en el que estos forman parte de la vida diaria de las personas (a veces sin que las mismas lo perciban). Esto ha influido en el interés por parte de la comunidad científica en estudiar a fondo dichos algoritmos, sus ventajas y limitaciones. La popularidad y la influencia en la vida diaria humana han llevado a varias preocupaciones ¿Hasta que punto podemos confiar en las decisiones tomadas por una máquina entrenada por un algoritmo de aprendizaje automático? ¿Cómo podemos saber que las decisiones que toma esta máquina son justas? Se puede entender como justicia o equidad a la ausencia de favoritismo o prejuicios hacia un individuo o grupo, basado en características inherentes o adquiridas.

Los algoritmos usan conjuntos de datos sacados de las sociedades humanas, por tanto, los sesgos existentes en estas pueden potencialmente verse reflejados en las decisiones tomadas por la máquina entrenada. Estos sesgos pueden, incluso, verse amplificados. Lo anterior deja ver la importancia que tiene para la sociedad el estudio de las formas de eliminación de los sesgos, ya sean, entre muchos otros, raciales o de género. La responsabilidad de esto cae en manos de los desarrolladores y el presente trabajo pretende dar ciertos aportes a la solución de dicho problema. Un ejemplo claro del riesgo de ignorar lo anterior es la popularización de representaciones semánticas pre-entrenadas en registros históricos, contenedores de sesgos sociales.

En los últimos años ha crecido el interés por el desarrollo de técnicas que permitan identificar y eliminar los sesgos en algoritmos de aprendizaje automático, pero que mantengan al mismo tiempo la precisión de los modelos.

Una de las principales desventajas del Aprendizaje de Máquina es el nivel de preparación que tiene que tener un investigador o ingeniero para sacar provecho de las técnicas en problemas reales. Para ello se han venido explorando muy recientemente alternativas que permitan automatizar el proceso de selección y entrenamiento de modelos de aprendizaje de máquina, permitiendo a los ingenieros e investigadores enfocarse en los detalles que no pueden ser automatizados o son de dominio muy específico y así atacar los problemas de una forma más ágil. A este tipo de técnicas se les conoce como técnicas de AutoML y están jugando un papel fundamental en la

democratización de la inteligencia artificial y el aprendizaje de máquinas.

Como parte de este trabajo se pretende utilizar una biblioteca de AutoML llamada AutoGoal, para asistir en el objetivo fundamental de obtener un método basado en ensembles que permita construir modelos con alta precisión para problemas de clasificación. Poniendo énfasis en la explotación de la diversidad entre los integrantes de los ensembles y permitiendo la optimización de mas de una métrica, abriendo paso a aplicaciones como la mitigación de los sesgos antes mencionados.

Para lograr el objetivo planteado, luego de una amplia curva de investigación y aprendizaje, se llegó a una solución que hace uso de algoritmos como Non-dominated Sorting y Probabilistic Grammatical Evolution.

Motivación

Un modelo de aprendizaje de maquina se entrena con el objetivo de optimizar una unica metrica, en la mayoría de los casos la precision. 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 estan desafortunadamente presente en la sociedad y por ende en los datos recopilados, en algunos casos incluso amplifican estos patrones negativos. Son varias la tecnicas 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 variacinoes en el metodo de entrenamiento con el mismo objetivo. Sin embargo permanece relativamente poco explorado el uso de tecnicas de optimizacion multiobjetivo que permitan al modelo optimizar hasta encontrar un buen balance entre cuan justo es y cuan 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 combinación de diferentes metricas en una sola, para poder aprovechar los multiples metodos de optimización que existen.

Capítulo 1

Estado del Arte

Capítulo 2

Propuesta

Capítulo 3

Detalles de Implementación y Experimentos

Conclusiones

Conclusiones

Recomendaciones

Recomendaciones