Czy gwałtowne upowszechnianie się rozwiązań opartych o metody uczenia maszynowego grozi dalszym pogłębianiem nierówności społecznych i nasileniem istniejących zjawisk dyskryminacyjnych - czy też może daje nadzieję na przyszłą redukcję tych problemów?

Szymon Rogus

Wstęp

Wraz z rozpowszechanieniem się sztucznej inteligencji oraz rozwiązań związanych z uczeniem maszynowym, wiele procesów w życiu codziennym zostało znacząco ułatwionych. Rozwój tej gałęzi informatyki znacząco przyspieszył proces automatyzacji wielu dziedzin naszego życia. Uczenie maszynowe jest swego rodzaju nowością, niepodobną do innych rozwiązań z zakresu informatyki. Oprócz wielu korzysci płynących z jego zastosowania, pojawiają się też pytania o pewne negatywne konsekwencje. Jedną z nich są właśnie nierówności społeczne oraz dyskryminacja. W tym eseju postaram się ustosunkować do obu aspektów.

Uczenie maszynowe a nierówności społeczne

Wydawawać by się mogło że uczenie maszynowe i dyskryminacja nie mają ze sobą nic wspólnego. Trzeba jednak pamiętać że wszystkie algorytmu, programy itd. Tworzone są przez ludzi. To sprawia że zdarzają się sytuacje takie jak ta opisana w artykule poniżej:

https://towardsdatascience.com/machine-learning-and-discrimination-2ed1a8b01038

Odnosi się on do oprogramowania opartego na uczeniu maszynowym **Correctional Offender Management Profiling for Alternative Sanctions** (**COMPAS**). W dużym skrócie – jest to algorytm, który jest wykorzystywany w sądownictwie w USA. Jego zadaniem jest przewidywanie prawodpodobieństwa tego że osoba oskarżona zostanie recydywistą. Artykuł wspomina o przydatności tego algorytmu. W wielu przypadkach jego działanie wspomogło pracę śledczych oraz policji. Algorytm wykorzystuje kwestionariusz złożony ze 137 czynników związanych z daną osobą, aby oszacować wynik.

Są niestety wspomniane sytuacje które negatywnie świadczą o **COMPAS**. Jest tam opisana sytuacja jaka miała miejsce w roku 2014. Po kilku kontrowersjach związanych z działaniem oprogramowania, prokurator generalny Eric Holder zasugerował na wykonanie badań skuteczności działaniu algorytmu. Badanie te (wykonane przez **ProPublica**) dały zaskakujące rezultaty. Według tej organizacji skutecznośc algorytmu była bardzo niska (ok 20%), a dalsze wyniki wykazały znaczące różnice w wynikach dla różnych ras. Algorytm

przeciętnie wskazywał że czarnoskórzy oskarżeni byli o 77 procent bardziej narażeni na ryzyko popełnienia w przyszłości brutalnego przestępstwa i o 45 procent bardziej prawdopodobne jes to, że popełnią w przyszłości jakiekolwiek przestępstwo.

Można zastanowić się nad tymi wynikami. Czy stanowią one dowód dyskryminacji, czy jest to zwyczajny wynik dla dostępnych danych? Tutaj można odnieść się do sposobu klasyfikacji. W przypadkach klasyfikacji dla dużych grup ludzi, często stosuje się pewne uproszczenia. Oznacza to że nie rozpatruje się indywidualnie każdego osoby, tylko często grupuje się ludzi według wieku, płci, rasy, wzrostu itd.

Takie grupowanie może prowadzić do pewnych zakłamań. Artykuł odnosi się do dwóch.

Odmienne traktowanie - polega na zaklasyfikowaniu kogoś w niedopuszczalny sposób. Obejmuje zamiar dyskryminacji, czego dowodem jest wyraźne odniesienie do przynależności do grupy.

Odmienny wpływ - analizuje konsekwencje klasyfikacji / podejmowania decyzji dla pewnych grup. Nie jest wymagana żadna intencja i jest z pozoru neutralna. Są to takie dane które wynikają bezpośrednio z faktu przynależności do jakiejś grupy.

Atrybut które są zagrożone dyskryminacją i działaniem czynników wyżej wymienionych to:

- rasa
- religia
- pochodzenie etniczne
- płeć
- niepełnosprawność
- wykonywany zawód
- poziom zamożności

Oba czynniki mogą powodować pewne przekłamania, przez co zachowanie COMPAS'u może wydawać się dyskryminujące w pewnych przypadkach. Wyobraźmy sobie sytuację w której algorytm przewiduje że osoby ze wzrostem powyżej 2m są wyjątkowo narażone na ponowne zostanie recydywistą oraz że popełnią w przyszłości przestępstwo. Czy oznacza to że musimy zachowywać specjalne srodki dla każdej osoby z tego grupy? Czy wzrost na pewno koreluje ze skłonnością do popełniania przestępstw?

Czy istnieje możliwość zmiany zachowania tego typu algorytmów? Przecież muszą używać jakichś danych. Artykuł sugeruje modyfikację danych uczących. Minimalizacja tzw. **Discriminatory Bias** (Dyskryminacyjnych danych), może prowadzić nie tylko do większej uczciwości ale też do poprawy działania algorytmu. Zasadniczym celem przy usuwaniu krytycznych danych jest to, aby pozostawić takie, które nie mają żadnych korelacji z wynikiem.

Co więcej, w przypadku takiego działania algorytmu tworzy się pewnego rodzaju zamknięte koło. Algorytm operując na takich danych, podejmuje nieuczciwe decyzje w wynikiem czego jest utrwalenie istniejących nierówności – te nowe dane będą wykorzysytwane przez podobne algorytmy w przyszłości. Takie sprzężenie zwrotne jest często bardzo głęboko ukryte w mechanizmach programów, więc o ile stwierdzenie że problem występuje, nie jest trudne, o tyle znalezienie źródła problemu i poprawne wyeliminowanie go jest znacznie cięższe.

Więc jakie są sposoby na wykrywanie które dane powoduję zaburzenia?

Problem z tymi danymi oraz ich usuwaniem lub zastępywaniem został ładnie opisany w tym artykule

https://hbr.org/2020/08/how-to-fight-discrimination-in-ai

Tekst odnosi się do tego z jakimi najcześciej aspketami dyskryminacyjnymi mają do czynienia firmy wdrażające tego typu rozwiązania. W artykule znowu jest mowa o **odmiennym traktowaniu** I **odmiennym wpływie** jako głównych czynnikach dyskryminujących. Warto zaznaczyc że jeśli chodzi o drugi czynnik – odmienny wpływ, to problem z nim jest dosyć złożony. W zależności od stopnia nierówności społecznych I systemowych, odmienny wpływ może być bardzo głęboko zakorzeniony w danych z pozoru neutralnych. Przykładowo, dla Polski powinniśmy użyć innych sposobów oraz oceny I analizy, które dane mimo pozornej neutralności wpływają na tworzenie się dyskryminacyjnych danych, niż zrobimy to dla USA czy Chin.

Jednym ze sposobów radzenia sobie ze zróżniocownym wpływem oraz sprawdzenia poziomu sprawiedliwości algorytmu (w pewien wybrany sposób) jest korzystanie z **reguły Pareta.** Zasada ta głosi że z 20% badanych obiektów związanych jest ok 80% pewnych zasobów. Zasada ta wykorzystywana jest m.in w obliczaniu wskaźnika Giniego (miara nierównomierności). W przypadku zróżnicowanego wpływu dzieli się ilość osób z grupy defaworyzowanej przez ilość osób z grupy faworyzowanej. Stosunek mniejszy niż 80% sugeruje istnienie odmiennego wpływu I niesprawiedliwe wyniki algorytmu. Nierównomierny rozkład danych wykorzysytwanych przy uczeniu na pewno będzie wpływał na wyniki.

Warto zauważyć także pozytywne strony takich tendencji. Występowanie takich problemów mobilizuje nas do podejmowania odpowienich kroków oraz tworzenia zapisów prawnych korygujących pewne procesy. Aktualnie prawodawstwo obejmuje także dane wykorzysytwane w Machine Learningu. Sam fakt takiego zjawiska zwiększa także naszą świadomość na istniejący problem.

Ciekawy aspekt poruszył poniższy artykuł:

http://sitn.hms.harvard.edu/uncategorized/2020/fairness-machine-learning/

Oprócz ponownego podniesienia tematu algorytmu COMPAS, jest tutaj wspomniany algorytm do wykrywania mrugnięć, wdrożony przez japońską firmę Nikon. Algorytm błędnie oceniał, ze Azjaci mrugają znacznie częściej niż osoby o innym pochodzeniu rasowym. Mimo, że firma nie zdradziła źródła problemu, to prawdopodobnie było to spowodowane zbyt jednorodnymi danymi – w procesie uczenia dane prawopodobnie w większości pokazywały twarze Azjatów (Jest to dobry przykład tego jak działą przeuczenie).

Powyższy przykład jest raczej przykładem błędu implementacyjnego, ale łatwo pomyśleć o analogii w algorytmach związanych z przestępczością, sądownictwem, udzielaniem pożyczek, wykrywaniem kłamstw itd. Zbyt jednorodne dane mogą przekłamywać rzeczywistość I stanowić odmienny wpływ.

Inny przykład to algorytmy wykorzystywane do wyświetlania reklam na Facebook'u. W zależności od tego co mamy w profilu (na podstawie naszego życia społecznego) wyświetlane są odpowiednie reklamy np. Ścieżek rozwoju kariery, ofery mieszkań, kredytów itd. Takie podejście może wydawać się słuszne, bo reklamy mają dostosowywać się do naszych zainteresowań I potrzeb, ale z drugiej strony pewna grupa ludzie nie dostanie szany którą dostanie inne grupa ludzi. Problem z równością szans jest wyjątkowo dotkliwy, ponieważ jest to kolejne ukryte sprzężenie zwrotne, które utrwala pewne negatywne stereotypy.

Zasadniczym problemem jest zatem sposób wybory danych. Nie można przecież zrównać ze sobą wszystkich grup społecznych, tak aby wyniki dla danego algorytmu były dopasowane do równych propocji w danej grupie. Trzeba pamiętać że celem dobrego algorytmu do przewidywania jest kombinacja skuteczności oraz sprawiedliwości.

A więc czy jest więcej sposobów na usunięcie krytycznych danych lub poprawę sposobu klasyfikacji?

Poniższy algorytm z perspektywy technicznej opisuję procesu krystalizowania uczciwych danych:

https://towardsdatascience.com/a-tutorial-on-fairness-in-machine-learning-3ff8ba1040cb

Do tej pory wymieniłem podejście eliminację danych wpływających na **odmienne traktowanie,** testowanie zbiory danych metodą Pareta oraz unikania używania zbyt jednorodnych zestawów danych.

Kolejnym wymienionym a artykule jest **Demographic Parity.** Jest to podstawowe kryterium tworzenia uczciwego zbiory danych. Polega na to w przybliżeniu na tym że wskaźniki akceptacji dla osób z dwóch różnych grup społecznych muszą być równe ze względu na przynależność do tych grup. Zapobiega to tworzeniu odmiennego wpływu. To kryterium często wykorzystuje się przy metodzie Pareta.

Przykład

Y – Wynik – Propozycja szkolenia z wybranej branży (1) lub brak propozycji (0)

A – Atrybut chroniony – Zamożny (1) lub biedny (0)

Skrótowo : $P(Y=1 \mid A=0) = P(Y=1 \mid A=1)$

Efektem jest parytet dla różnych grup społecznych jeśli chodzi o szanse.

Equalized odds to kolejna metoda do usprawnianie zestawu danych. Zakłada niezależność predyktora I wrażliwego atrybut, w zależności od wyniku.

Przykład:

R - Predyktor – czy osoba była uznana za wybitnego ucznia (1) czy też nie (0)

Y - Wynik – dostanie się na elitarne studia za granicą (1) lub nie (0)

A - Atrybut chroniony – Osoba jest biała (1) lub czarna (0)

Metoda zakłada niezależność atrubutu (koloru skóry) oraz wyników w nauce, dla osób które zostały przyjęte na elitarne studia za granicą. Oznacza to że dążymy do sytuacji, której dla algorytmu liczy się tylko fakt bycia wybitnym uczniem.

Skrótowo: $P(R=1|A=0, Y=1) = P(R=1|A=1, Y=1)$.
Well-calibrated systems zakłada że nie występuje zależność między wynikiem a atrybutem chronionym, ale istnieje zależność między wynikiem a predyktorem

Przykład:

R - Predyktor – czy osoba była uznana za wybitnego ucznia (1) czy też nie (0)

Y - Wynik – dostanie się na elitarne studia za granicą (1) lub nie (0)

A - Atrybut chroniony – Osoba jest biała (1) lub czarna (0)

Na pierwszy rzut oka metoda ta jest identyczna jak poprzednia (**Equalized odds**), różnią się jednak kolejnością wnioskowania

Skrótowo: $P(Y=1|A=0, R=1) = P(Y=1|A=1, R=1)$.

Tutaj na podstawie tego że osoba jest czarna lub biała nie zmienia się jej prawdopodobieństwa na elitarne studia za granicą. Dla **Equalized odds** na podstawie tego że osoba dostała się na te studia, nie możemy stwierdzić prawdopodobieństwa jej koloru skóry. Różnica między tymi dwoma metodami jest nieduża.

Ogólnie zakłada się że rozwiązanie uczenia maszynowego jest sprawiedliwe jeśli spełnia te trzy warunki.

Dosyć istotnym przykładem dużego rozwiązania, które zostało uznane za dyskryminujące, obok COMPAS'u oraz algorytmu w firmie Nikon, jest algorytm zatrudniania firmy Amazon. W 2015 firma zadała sobie sprawę (po wielu uwagach na ten temat), że ich algorytm faworyzował mężczyzn. Jednym z głównych czynników, było analizowanie przez program ilości podań o pracę wysłanych do firmy na przestrzeni ostatnich 10 lat. Większość aplikatnów to mężczyźni, więc statystycznie dla dowolnej próby dwójki aplikantów – mężczyzny I kobiety – program korzystniej oceniał mężczyzne (a przynajmniej na korzystniejszą ocenę wpływał ten konkretny czynnik). Jest to idealny przykład złamania wszystkich trzech wyżej wymienionych zasad:

- Brak parytetu szansy ze względu na płeć - brak **Demographic Parity**
- Zależność atrybutu chronionego (płeć) od wyniku zatrudnienie (**Equalized odds** oraz **Well-calibrated systems**)

Widać także w tym przypadku nie tylko **odmienny wpływ,** ale także **odmienne traktowanie.** Celowe dodanie do algorytmu czynnika związanego z stosunkiem płci w podaniach o pracę na przestrzeni danego okresu jest świadomym czynnikiem dyskryminacji.

Nawet w przypadku ML zdarza się też dyskryminowanie systemowe. W 2019 roku, magazyn Sciene opublikował artykuł (https://science.sciencemag.org/content/366/6464/447) o algorytmie stosowanym przez amerykańską służbę zdrowia. Algorytm miał za zadanie oceniać na podstawie wielu czynników, czy pacjent będzie potrzebował dodatkowej opieki medycznej na bardziej specjalistycznym poziomie. Algorytm faworyzował ludzi o rasie białej dosyć znacząco. Sam fakt przynależności rasowej nie był zawarty jako atrybut w danych uczących, ale użyte zostały innych atrybut mocno skorelowane z rasą, takie jak:

- Historia hosztów opieki zdrowotnej pacjenta
- Przychody pacjenta
- Średnia długość życia

Po raz kolejny takie zachowanie algorytmu tworzy zamknięte koło. Częstsze odrzucanie osób czarnoskórych przez algorytm, pogłębia ten podział, przez co tworzy nowe niepoprawne I niesprawiedliwe dane.

Podsumowanie

Problem dyskryminacji w uczeniu maszynowym, nie jest problemem który powinniśmy bagatelizować. Algorytmy przyjmują dane takie, jakie dostarczy im człowiek więc nie można powiedzieć by były dsykryminujące. To my je tworzymy, a więc możemy zadbać o sprawidliwości danych. Mimo wielu wymienionych problemów, zdecydowanie warto zauważyć także plusy. Jesteśmy świadomi istniejącego problemu I podejmujemy kroki aby mu przeciwdziałać. Składają się na nie zarówno kroki prawne jak I zwiększanie świadomości w tym zakresie oraz poprawnianie rozwiązań które pogłębiają nierówności. Paradoksalnie, fakt występowania tych problemów może sprawić, że nierówności I dyskryminacja zaczną się zmiejszać. Ciągła poprawa obecnych rozwiązań sprawi że, problem może być w przyszłości mniej znaczący. Trzeba wziąć też pod uwagę fakt, że programy I algorytmy oparte na uczeniu maszynowym popełaniają błędy, ale w przeciwnieństiwe do ludzi nie kierują się uprzedzeniami – są obiektywne. To czy tak się stanie zależy tylko od nas.