Strona czasopisma: http://analit.agh.edu.pl/

Techniki powiększania danych stosowane w uczeniu maszynowym

Data Augmentation Techniques used in Machine Learning

Justyna Ścigaj

AGH Akademia Górniczo-Hutnicza, Wydział Inżynierii Materiałowej i Ceramiki, al. Mickiewicza 30, 30-059 Kraków

ABSTRAKT: Celem niniejszego referatu jest przybliżenie technik powiększania danych stosowanych w uczeniu maszynowym. Dzięki generowaniu nowych danych, na podstawie istniejących można w sposób łatwy uzyskać nowe zdjęcia, teksty, a także dźwięki. W jaki sposób to się dzieje, a także jakie są dostępne techniki, opisano w poniższym artykule. Przedstawiono również wady i zalety technik.

ABSTRACT: The aim of this paper is to present data augmentation techniques used in machine learning. By generating new data, you can easily obtain new photos, texts and sounds based on the existing ones. How this happens, as well as what techniques are available, are described in the article below. Advantages and disadvantages of the techniques are also presented.

Słowa kluczowe: uczenie maszynowe, data augmentation, SMOTE

1. Wstęp

Powiększanie danych to nic innego jak proces sztucznego generowania nowych danych, na podstawie już istniejących. Wykorzystywany ten proces jest najczęściej wtedy, kiedy chcemy powiększyć nasz zbiór danych, ale wiąże się to ze znacznymi kosztami, przez co zastosować możemy tzw. argumentację danych.

Technika ta znalazła swoje zastosowanie w analizie obrazu. Dlaczego? Ponieważ obrazy są bardzo podatne na drobne modyfikacje. Dla ludzkiego oka zdjęcie będzie wyglądało na takie samo, dla algorytmu jest zmienione. Najprościej wytłumaczyć to na przykładzie samochodu. Możemy stać przed nim i patrzeć na niego centralnie albo też nieco z boku. Dla naszego mózgu to będzie wciąż ten sam pojazd, jednak dla algorytmu to zupełnie inny obiekt, z innej perspektywy.

Co tak naprawdę możemy zrobić z obrazem, który chcemy sztucznie przetworzyć? Rozwiązań jest nieskończenie wiele. Obraz możemy lekko obrócić, w dowolnym kierunku, a także o dowolny kąt. Możemy przesunąć w lewo, w prawo, w górę i w dół. Zmienić jego kolorystykę, czy także dokonać mniej lub więcej zmian, np. zmiana jasności, kontrastu, które dadzą mnóstwo informacji modelowych danych.

Istotne jednak jest, żeby nasz model był poprawnie generalizowany, czyli żeby skutecznie sobie radził z nowymi, nieznanymi danymi, spoza zbioru uczącego. Najprościej zrozumieć to na przykładzie klasyfikacji obrazów. Mamy model, który został wytrenowany na danych, w celu rozpoznawania kotów i psów na zdjęciach. Jeżeli został on poprawnie zoptymalizowany i dobrze uogólnia wyuczone wzorce, to nawet jeśli dostanie on nowe zdjęcia kotów i psów, których wcześniej nie widział, to również poprawnie je zaklasyfikuje. A jeśli nie będzie poprawnie generalizował to wtedy może mieć problem z rozpoznaniem nowych obrazów kotów i psów [1].

2. SMOTE – SYNTHETIC MINORITY OVER-SAMPLING TECHNIQUE

SMOTE jest jedną z zaawansowanych technik powiększania zbioru danych, która znajduje szerokie zastosowanie w uczeniu maszynowym. Działa ona na zasadzie generowania przykładów danych syntetycznych dla klasy mniejszości poprzez interpolację między sąsiadami [8,9].

Ogólny proces SMOTE można opisać w następujący sposób:

a) Wybór próbek: Najpierw wybierz próbki z klasy mniejszości, które mają zostać rozszerzone.
b) Znajdowanie sąsiadów: Dla każdej wybranej próbki algorytm znajduje k najbliższych sąsiadów w przestrzeni cech. Na tym etapie trzeba zdecydować, jakie wartości k wybrać. Zwykle używana jest wartość 5.
c) Generowanie przykładów syntetycznych: Dla każdej wybranej próbki algorytm wybiera jednego z k sąsiadów i tworzy nową próbkę poprzez interpolację wartości cech pomiędzy nimi.
d) Powtarzanie procesu: Proces generowania próbek syntetycznych powtarza się w celu zwiększenia liczby próbek klas mniejszościowych [8,9].

SMOTE jest użyteczną techniką radzenia sobie z problemem niezrównoważonych klas w danych treningowych, ale należy ją stosować świadomie, dostosowując parametry do konkretnego problemu i danych treningowych.

Zalety SMOTE:

- Zwiększa równowagę klas – SMOTE pomaga zrównoważyć proporcje klas w danych szkoleniowych, generując syntetyczne przykłady, co może prowadzić do zwiększonego zapotrzebowania na zasoby obliczeniowe i pamięciowe
- Unikanie nadmiernego dopasowania – przy odpowiednich parametrach SMOTE może wprowadzać większą różnorodność do danych uczących i utrudnić modelowi zapamiętywanie konkretnych przypadków
- Prosta implementacja – Technika ta jest stosunkowo prosta do zrozumienia i zastosowania [8,9]

Wady SMOTE:

- Zwiększanie rozmiaru zbioru danych – SMOTE może znacznie zwiększyć rozmiar zbioru uczącego poprzez generowanie syntetycznych przykładów, co może prowadzić do zwiększonego zapotrzebowania na zasoby obliczeniowe i pamięciowe
- Możliwość zniekształcenia danych – niewłaściwe użycie SMOTE lub niewłaściwy dobór parametrów może prowadzić do zniekształcenia danych uczących, co może obniżyć wydajności modelu
- Nie uwzględnia struktury danych – SMOTE generuje syntetyczne przykłady poprzez interpolację pomiędzy istniejącymi przykładami, co może prowadzić do utraty struktury lub wprowadzenia sztucznych wzorców
- Wrażliwość na szum – SMOTE może być wrażliwy na szum w danych, co oznacza, że wygenerowane przykłady mogę być niedokładne lub nieprawdziwe, co może prowadzić do pogorszenia wydajności modelu
- Potrzeba dobrego doboru parametrów – wybór właściwych parametrów, takich jak liczba sąsiadów i liczba wygenerowanych przykładów ma kluczowe znaczenie dla efektywności SMOTE [8,9]

3. FEATURE-BASED AUGMENTATION

To technika powiększania zbioru danych, która polega na modyfikowaniu funkcji lub atrybutów danych szkoleniowych, w celu wygenerowania nowych przykładów. Pozwala to zwiększyć różnorodność danych szkoleniowych i poprawić wydajność modelu uczenia maszynowego.

Dzięki tej technice można zmieniać jasność i kontrast obrazów, dodawać szum do danych dźwiękowych lub dodawać synonimów do danych tekstowych. Feature-based augmentation jest techniką uniwersalną i można ją zastosować do różnych dziedzin i problemów, a nie tylko problemów klasyfikacyjnych [10,11].

Ogólny proces można opisać w następujący sposób:

a) Wybór odpowiednich cech do modyfikacji: Pierwszym krokiem do stosowania FBA jest identyfikacja cech danych, które można modyfikować w celu wygenerowania nowych przykładów. Wybór odpowiednich funkcji zależy od rodzaju danych i problemu, jaki chcemy rozwiązać. Przykładowo w przypadku danych obrazu mogą to być takie cechy jak jasność, kontrast, kąt obrotu czy przesunięcie.
b) Definicja sposobów modyfikowania cech: Następnie należy zdefiniować, w jaki sposób chcemy modyfikować wybrane cechy. Przykładowo w przypadku danych obrazowych możemy zmienić jasność zwiększając lub zmniejszając wartość piksela, zmieniać kontrast dostosowując różnicę między jasnymi i ciemnymi obszarami czy obrócić obraz o określony kąt.
c) Generowanie nowych przykładów: Po zdefiniowaniu modyfikacji funkcji można rozpocząć generowanie nowych przykładów, stosując wybrane modyfikacje do istniejących danych.
d) Zastosowanie w uczeniu maszynowym: wygenerowane nowe przykłady danych można wykorzystać w procesie uczenia maszynowego razem z oryginalnymi danymi szkoleniowymi. Daje to modelowi uczenia maszynowego więcej przykładów do nauczenia, co może pomóc poprawić jego wydajność i zdolność do generalizacji [10,11]

Zalety FBA:

- Ograniczona zależność od dostępności danych – FBA można stosować nawet przy ograniczonej liczbie danych szkoleniowych, ponieważ wymaga modyfikacji istniejących funkcji, a nie generowania zupełnie nowych danych. Pozwala to na zwiększenie różnorodności zbioru danych nawet przy ograniczonym dostępie do danych treningowych
- Zachowanie struktury danych – FBA pozwala zachować strukturę i kontekst danych treningowych, ponieważ opiera się na modyfikacji istniejących funkcji. Dzięki temu zmiany wprowadzone przez FBA są zazwyczaj subtelne i nie powodują zakłóceń danych
- Szybkość w porównaniu do innych technik powiększania danych
- Elastyczność – FBA można zastosować w różnych domenach uczenia maszynowego [10,11]

Wady FBA:

- Ryzyko utraty ważnych informacji - Modyfikacje funkcji wprowadzone przez FBA mogą prowadzić do utraty lub zmiany ważnych informacji w danych treningowych
- Możliwość wprowadzenia szumu – niekontrolowane modyfikacje funkcji mogą wprowadzić szum lub zniekształcenia, które mogą pogorszyć wydajność modelu
- Zależność od jakości danych wejściowych – Skuteczność FBA może być ograniczone jakością danych wejściowych i jakością zastosowania modyfikacji funkcji. W przypadku danych niskiej jakości modyfikacja cech może nie przynieść oczekiwanych rezultatów
- Potencjalna potrzeba dostosowywania – wybór odpowiednich modyfikacji funkcji i dostosowanie ich do konkretnego problemu może wymagać eksperymentowania i dostosowywania, co może być czasochłonne [10,11]

4. Techniki stosowane w NATURAL LANGUAGE PROCESSING

To technika powiększania zbioru danych stosowana w przetwarzaniu języka naturalnego, która polega na generowaniu nowych przykładów tekstowych poprzez modyfikację istniejących danych. Celem jest zwiększenie różnorodności szkoleniowych, co może poprawiać wydajność modeli NLP w różnych zadaniach, takich jak klasyfikacja tekstu, analiza tonacji i generowanie tekstu [2,7].

Ogólny proces można opisać w następujący sposób:

a) Zmiany w strukturze tekstu: Jednym ze sposobów poszerzenia danych jest modyfikacja struktury tekstu, na przykład poprzez zmianę kolejności słów, dodanie lub usunięcie słów lub przekształcenia zdania w pytanie. Zmiany te mogą wprowadzić różnorodność danych szkoleniowych
b) Zastępowanie słów synonimami: innym popularnym podejściem jest zastępowanie słów w zdaniach ich synonimami. Można to zrobić za pomocą słowników synonimów lub modeli językowych, które mogą generować synonimy dla danego słowa
c) Dodawanie szumu do tekstu: np. poprzez wprowadzenie literówek, błędów ortograficznych lub gramatycznych, może pomóc w zwiększaniu różnorodności danych i poprawie odporności modelu na tego typu szumy na wejściu
d) Generowanie nowych przykładów: powiększanie danych o nowe przykłady tekstowe, które można wykorzystać w procesie uczenia modeli NLP [7]

Zalety NLP

- Zwiększanie różnorodności danych – większa wydajność modeli NLP w rozpoznawaniu różnych wzorców szkoleniowych
- Ulepszona generalizacja – dzięki zwiększonej różnorodności modele NLP mogą lepiej generalizować na nowe, nieznane dane, co może prowadzić do lepszych wyników w praktyce
- Ograniczenie nadmiernego dopasowywania – poprzez wprowadzenie różnorodności
- Elastyczność – techniki NLP są stosunkowo elastyczne i można je dopasowywać do różnych typów danych tekstowych [2,7]

Wady NLP

- Możliwość wprowadzenia szumu – niekontrolowane dodanie szumu lub modyfikacja może wprowadzić niechciany szum
- Ryzyko utraty znaczenia – możliwość utraty znaczenia tekstu pierwotnego
- Zależność od jakości danych wejściowych – jeśli dane wejściowe są niskiej jakości, dane wygenerowane w wyniku augmentacji również mogą być nieskuteczne
- Zwiększony wysiłek obliczeniowy – Generowanie wielu wariantów danych tekstowych może wymagać znacznego wysiłku obliczeniowego [2,7]

5. Inne techniki powiększania danych

1) Geometric Data Augmentation – Powiększanie danych metodami geometrii obliczeniowej
2) Generative Data Augmentation – Powiększanie danych za pomocą modeli generatywnych
3) Audio Data Augmentation – Powiększanie danych w zakresie przetwarzania dźwięku [1]

Literatura

[1] https://aigeekprogrammer.com/pl/sieci-konwolucyjne-data-augmentation/
[2] https://saturncloud.io/glossary/data-augmentation-in-natural-language-processing-nlp/
[3] https://en.wikipedia.org/wiki/Data_augmentation#Synthetic_oversampling_techniques_for_traditional_machine_learning
[4] https://www.blog.trainindata.com/oversampling-techniques-for-imbalanced-data/
[5] https://ietresearch.onlinelibrary.wiley.com/doi/full/10.1049/cit2.12123
[6] https://www.kaggle.com/code/marcinrutecki/smote-and-tomek-links-for-imbalanced-data
[7] https://www.ibm.com/topics/natural-language-processing
[8] https://learn.microsoft.com/en-us/azure/machine-learning/component-reference/smote?view=azureml-api-2
[9] https://medium.com/@corymaklin/synthetic-minority-over-sampling-technique-smote-7d419696b88c
[10] https://www.sciencedirect.com/science/article/abs/pii/S156625352300091X

[11] https://arxiv.org/abs/2007.08505