Wydział Elektroniki
Politechnika Wrocławska

Analiza porównawcza wybranych metod selekcji cech w zadaniu klasyfikacji danych niezbalansowanych

Maciej Hajduk
Nr indeksu: 236596

Praca magisterska napisana
pod kierunkiem
dr Pawła Trajdosa

Politechnika
Wrocławska
Wrocław 2021

Spis treści

1 Wstęp

1.1 Wprowadzenie

Celem pracy jest porównanie różnych metod selekcji cech w problemie trenowania algorytmów uczenia maszynowego na danych niezbalansowanych. W jej ramach, przedstawione i opisane zostaną popularne obecnie metody selekcji oraz przeprowadzone zostaną eksperymenty dla przykadowych zbiorów danych, zarówno naturalnych jak i sztucznie wygenerowanych, celem których będzie stworzenie rankingu algorytmów. Autor sprawdzi, jak właściwie przeprowadzona selekcja wpływa na jakość wyników dostarczanych przez klasyfikator i jak przytoczone przez niego metody radzą sobie z danymi, w których występuje znaczna przewaga liczebności jednej bądź kilku klas. Aspekt inżynieryjny opierać się będzie na implementacji zaproponowanych w pracy eksperymentów, co pozwoli na kompleksowe porównanie algorytmów..

Praca swoim zakresem obejmie eksperymenty…

Praca składa się z sześciu rozdziałów:

Rozdział pierwszy: Omówienie analizy wybranego problemu, przedstawienie motywacji podjęcia tego tematu oraz…

Rozdział drugi: Przegląd literaturowy…

Rozdział trzeci: Szczegółowa charakterystyka zagadnienia…

Rozdział czwarty: Założenia i plan eksperymentu…

Rozdział piąty: Wyniki eksperymentów…

Rozdział szósty: Podsumowanie otrzymanych wyników, wnioski…

Udało się zrealizować wszystkie postawione cele.

2 Analiza problemu

W tym rozdziale przedstawiona będzie analiza problemu...

Uczenie maszynowe to bardzo dynamicznie rozwijająca się gałąź informatyki. Niezwykła ekspansja wynika z zapotrzebowania na wykrywanie prawidłowości, uogólnianie oraz precyzowanie danych. Takie możliwości pozwoliły znaleźć zastosowanie dla algorytmów sztucznej inteligencji w bardzo wielu różnych branżach - począwszy od medycyny, poprzez finanse, produkcję i branżę rozrywkową. Tak duży przekrój różnych zastosowań wymaga ciągłego ulepszania istniejących już wzorców oraz wymyślania nowych, lepszych i bardziej efektywnych algorytmów. Jednym z podstawowych etapów każdego programu bazującego na uczeniu maszynowym jest selekcja cech. Jest odpowiedzialna za wybór najbardziej istotnych atrybutów badanych obiektów, co przekłada się bezpośrednio na poprawne działanie klasyfikatora. Dyspozycja coraz większymi bazami danych zmusza do optymalizacji tego procesu. Gwałtownie rosnąca liczba cech stanowi poważny problem - powoduje nie tylko wydłużenie procesu uczenia oraz wzrost złożoności klasyfikatora, ale niesie ze sobą także ryzyko spadku poprawnej klasyfikacji. Związane jest to z tak zwanym ”przekleństwem wymiarowości”. Zjawisko to zachodzi, gdy ilość cech znacznie przewyższa liczebność samego zbioru danych. Zadaniem selekcji cech jest również lepsze zrozumienie problemu oraz zmniejszenie kosztów archiwizacji przyszłych danych. W kolejnych rozdziałach opisane zostaną trzy główne metody tworzenia algorytmów selekcji: metody rankingowe - zwane filtrami, metody opakowane oraz metody wbudowane. Dla każdej z wymienionych metod zostanie określona idea, oraz przedstawione zostaną algorytmy reprezentujące daną metodologie.

3 Przegląd literaturowy

W tym rozdziale zostaną pokrótce opisane wybrane artykuły naukowe zajmujące się tematyką selekcji cech. Jest to, szczególnie ostatnio, często poruszany 1problem, co skutkuje dużym przekrojem prac, również w ujęciu czysto dziedzinowym - jak wykorzystanie konkretnych algorytmów dla bardzo konkretnych zastosowań.

Wstępną analizę problemu proponuje Jakub Piątkowski w pracy *Analiza i rozwój metod selekcji cech dla dużych problemów klasyfikacyjnych*. Autor opisuje w pracy istniejące już, popularne metody selekcji cech z dokładnym opisem każdej z nich. Jest to dobry wstęp do samego problemu klasyfikacji i selekcji cech, każda z metod posiada również opisane wzorami zaplecze matematyczne. W podobnej pracy - *Redukcja wymiarowości i selekcja cech w zadaniach klasyfikacji i regresji z wykorzystaniem uczenia maszynowego*, twórca przytacza wyniki przeprowadzonych przez siebie eksperymentów, co pozwala mu na stworzenie rankingu algorytmów. Tezy postawione przez autora zostaną sprawdzone w niniejszej pracy. Bardzo szczegółowo do zagadnienia podszedł mgr Wiesław Chmielnicki, pisząc rozprawę *Efektywne metody selekcji cech i rozwiązywania problemu wieloklasowego w nadzorowanej klasyfikacji danych*. Poza opisem znanych metod, rozprawa zawiera propozycje nowych algorytmów hybrydowych, które pozwoliły uzyskać w niektórych przypadkach lepsze efekty, niż najpopularniejsze, używane zazwyczaj. Interesujący temat poruszył Mark A. Hall w swojej książe *Correlation-based Feature Selection for Machine Learning*, badając wartość zestawu cech na podstawie korelacji pomiędzy nimi. Autor prowadzi szereg eksperymentów porównując swoje metody do metod stosowanych ogólnie, starając się między innymi wyodrębnić problemy, dla których jego algorytm jest najbardziej skuteczny. Twórcy pracy *A Survey on Evolutionary Computation Approaches to Feature Selection* zajęli się przeglądem znanych metod tworząc dokument podsumowujący każdą z nich, z jej wadami oraz zaletami. Artykuł jest oparty na przeglądzie najnowszych prac w zadanej dziedzinie i pozwala na dobranie odpowiedniej metody do zadanego zadania. Ważna dla tematu niniejszej pracy jest również rozprawa *Zastosowanie wybranych metod przekształcania i selekcji danych oraz konstrukcji cech w zadaniach klasyfikacji i klasteryzacji* mgr inż. Piotra Płońskiego [**?**], podsumowująca niejako cały proces uczenia maszynowego i roli, jaką pełni selekcja cech w kontekscie dobrze działającego systemu analizy danych.

4 Szczegółowa charakterystyka zagadnienia

W tym rozdziale omówiona zostanie szczegółowa analiza zagadnienia.

4.1 Metody Rankingowe

Najprostsze podejście do problemu selekcji cech reprezentowane jest właśnie poprzez metody rankingowe, nazywane też filtrami. Jak sama nazwa wskazuje do zadania selekcji przy pomocy metod rankingowych podchodzimy wyróżniając w zbiorze cech następujące grupy: cechy istotne, nieistotne i redundantne. Istotne to takie cechy, które odróżniają od siebie klasy, nieistotne to takie których wartości dla problemu klasyfikacji są przypadkowe, a cechy redundantne to takie których role z powodzeniem mogą przyjąć inne cechy. Metody rankingowe polegają więc na znalezieniu pewnej miary pozwalającej stworzyć taki ranking cech, a potem wybrać najlepsze cechy, a odrzucić najgorsze. Metody rankingowe zazwyczaj są najszybsze i - co istotne - nie zależą one od używanej metody analizy danych. Ich istotną wadą stanowi brak możliwości uwzględnienia zależności pomiędzy cechami. Kolejne opisane typy metod selekcji cech tej wady nie posiadają.

4.2 Metody opakowane

Podstawowymi metodami selekcji cech są metody opakowane, tak zwane wrappery. W przeciwieństwie do metod rankingowych, w których selekcja cech i klasyfikator pozostają niezależne, w algorytmach opakowanych selekcji, ocena atrybutów dokonuje się przy użyciu konkretnego modelu. To właśnie efektywność samego klasyfikatora służy za miarę skuteczności metody. Zaletą tej metody jest jej uniwersalność i dokładność, natomiast wadą - wysoka złożoność obliczeniowa. Dla efektywności tych algorytmów istotny jest sposób ustalania podzbioru cech. Wśród wielu metod wyszukiwania tegoż, wyróżnić można najprostszą - przeszukanie całego zbioru podzbiorów. Jest to jednak rozwiązanie bardzo kosztowne. Wobec tego typowymi strategiami są: przeszukiwanie w przód, przeszukiwanie wstecz oraz tworzenie indywidualnego rankingu.

4.3 Metody wbudowane

Metody wbudowane zawierają się w algorytmie klasyfikacji i to na etapie tworzenia modelu przypisuje się poszczególnym cechom wagi lub przeprowadza się ich eliminację. Do algorytmów klasyfikacji z wbudowaną metodą selekcji zaliczyć można popularne LASSO i RIDGE. W literaturze natknąć się też można na przypasowanie do tej kategorii metody wektorów nośnych (SVM) czy też analizy składowych głównych (PCA). Zaletą tych metod jest ich szybkość, ponieważ użycie ich nie wiąże się z dodatkowymi operacjami na zbiorze.

5 Założenia i plan eksperymentu

Temat projektu zakłada przeprowadzenie serii eksperymentów porównujących efektywność popularnych metod selekcji cech. Bazy danych użyte do eksperymentów zakładają skupienie się na problemach wieloklasowych, których elementy są opisywane przez dużą ilość cech. Hipoteza, z którą twórcy będą konfrontować wyniki eksperymentów to założenie, że wszystkie metody selekcji cech poradzą sobie podobnie z postawionym zadaniem i poza względami wydajnościowymi, nie ma znaczenia funkcja, która zostanie użyta. W celu sprawdzenia hipotezy, postanowiono skorzystać skorzystać zarówno z gotowych rozwiązań pozwalających na generowanie syntetycznych problemów klasyfikacyjnych. Badania zostaną przeprowadzone z pomocą języka Python w wersji 3.8 oraz bibliotek: pandas, sklern i numpy. Sposób generowania danych opisany został w rozdziale Generowanie danych. Kolejne eksperymenty zostały opisane poniżej..

6 Wyniki

7 Wnioski

8 Bibliografia

[1] Feature Selection for High-Dimensional and Imbalanced Biomedical Data Based on Robust Correlation Based Redundancy and Binary Grasshopper Optimization Algorithm, Garba Abdulrauf Sharifai and Zurinahni Zaino; 2020.

[2] A Survey on Evolutionary Computation Approaches to Feature Selection, Bing Xue; Mengjie Zhang; Will N. Browne; Xin Yao, 2015.

[3] Correlation-based Feature Selection for Machine Learning, Mark A Hall, University of Waikato, 1999.

[4] Efektywne metody selekcji cech i rozwiazywania problemu wieloklasowego w nadzorowanej klasyfikacji danych, Wieszław Chmielnicki, IPPT PAN, 2012

[6] Zastosowanie wybranych metod przekształcania i selekcji danych oraz konstrukcji cech w zadaniach klasyfikacji i klasteryzacji, Piotr Płoński, Politechnika Warszawka, 2016.

[7] Analiza i rozwój metod selekcji cech dla dużych problemów klasyfikacyjnych, Jakub Piątkowski, Uniwersytet Mikołaja Kopernika, 2006.

[8] Feature Selection in Imbalance data sets, International Journal of Computer Science Issues, 2013.

9 Zawartość płyty CD

Do pracy dołączono płytę CD o następującej zawartości:

- kod źródłowy programu znajdujący się w folderze `/src`
- gotową, zbudowaną w katalogu `/dist` aplikację
- katalog `/docs` zawierający kod źródłowy tej pracy
- plik w formacie `pdf` zawierający pracę