

Sprawozdanie - Pracownia Projektowa

Pavlo Filipchuk, Anna Bujak, Maciej Jakubiak

12 czerwca 2015

Spis treści

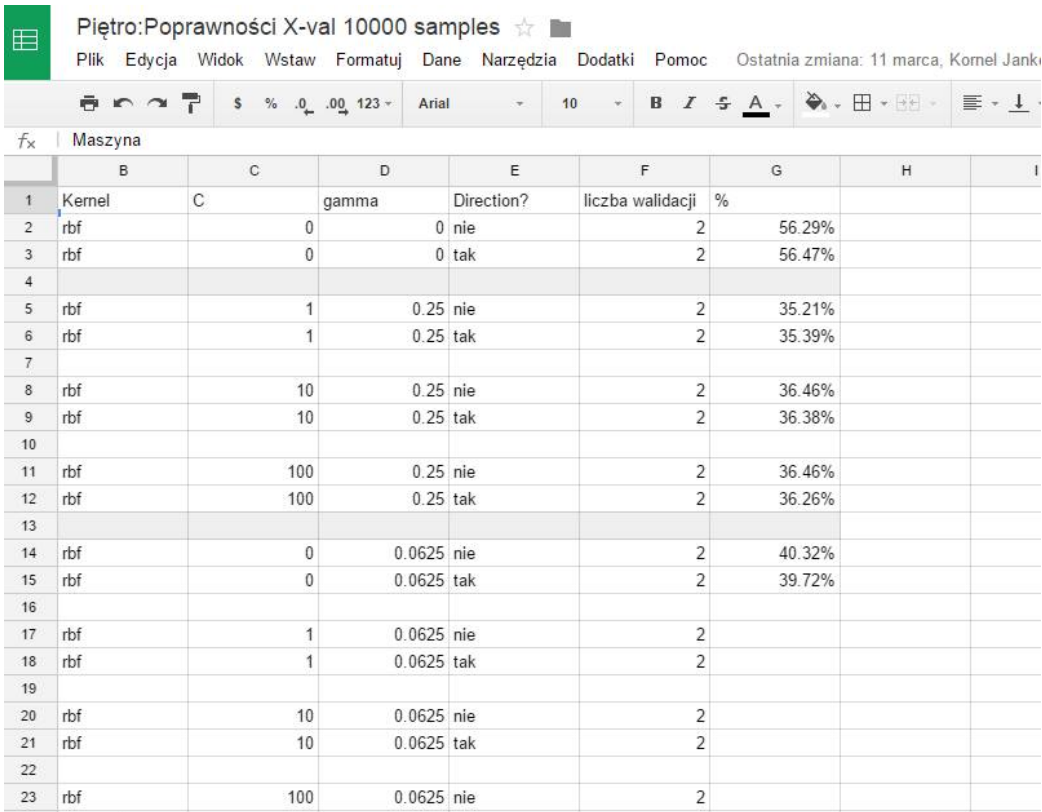
1	Wstęp	2
2	Pierwsze próby	3
3	Praca na danych z MATLABem	3
3.1	Wycięcie prostopadłościanu z danych	4
3.2	Wartościowanie anten	4
4	Szukanie klasyfikatora piętra	6
5	Szukanie aproksymatora Y	7
6	Selekcja cech	7
7	Zakończenie	7

1 Wstęp

Mieliśmy do rozwiązania problem, który polegał na tym, że po otrzymaniu sygnału z anten sieci komórkowej musieliśmy zlokalizować dany telefon w budynku wydziału MiNI. Do określenia były współrzędne x, y oraz piętro w budynku.

Projekt ten wykonywaliśmy w ośmioosobowej grupie. Co tydzień spotykaliśmy się na zajęciach, gdzie omawialiśmy postępy w zadaniu i stawialiśmy sobie nowe cele, zadania, a także rozpatrywaliśmy potencjalne problemy. Stworzyliśmy także grupę dyskusyjną, gdzie omawialiśmy rezultaty działań i zawieraliśmy istotne spostrzeżenia nt. projektu. Powstała również wspólna przestrzeń dyskowa, gdzie udostępnialiśmy sobie nawzajem różne dane, skrypty, wyniki, informacje, dokumenty, poradniki, wykresy i statystyki.

Do próby rozwiązania problemu wykorzystaliśmy uczenie maszynowe. Używaliśmy oprogramowania RapidMiner.



	B	C	D	E	F	G	H	I
1	Kernel		gamma	Direction?	liczba walidacji	%		
2	rbf	0	0	nie	2	56.29%		
3	rbf	0	0	tak	2	56.47%		
4								
5	rbf	1	0.25	nie	2	35.21%		
6	rbf	1	0.25	tak	2	35.39%		
7								
8	rbf	10	0.25	nie	2	36.46%		
9	rbf	10	0.25	tak	2	36.38%		
10								
11	rbf	100	0.25	nie	2	36.46%		
12	rbf	100	0.25	tak	2	36.26%		
13								
14	rbf	0	0.0625	nie	2	40.32%		
15	rbf	0	0.0625	tak	2	39.72%		
16								
17	rbf	1	0.0625	nie	2			
18	rbf	1	0.0625	tak	2			
19								
20	rbf	10	0.0625	nie	2			
21	rbf	10	0.0625	tak	2			
22								
23	rbf	100	0.0625	nie	2			

Pomocny również okazał się program MATLAB, w którym pisaliśmy po-

mocne skrypty takie jak:

- Generator trójwymiarowych map, które pokazywały rozkładanie się błędu na współrzędnych x i y
- Generator dwójwymiarowych map siły sygnału z anteny
- 'Wycięcie' prostopadłościanu danych
- Wartościowanie anten

Dokładniejszy opis powyższych skryptów został zamieszczony w rozdziale trzecim.

2 Pierwsze próby

Za cel postawiliśmy sobie uzyskanie klasyfikatora, który dla wektora pomiarów ustala piętro, na którym mogły zostać wykonane.

Wykonaliśmy wstępną analizę danych, zbadaliśmy osiągi różnych klasyfikatorów dla pierwotnej postaci danych. Następnie okazało się, że należy na nie nanieść różne poprawki, niezbędne do dalszych badań. Szczególny problem okazał się z niefortunnie dobraną wartością -1 jako oznaczeniem braku sygnału z danej anteny. Wartość tą zmienialiśmy na -200 (próby użycia innych wartości dawały nielepsze rezultaty), ponieważ siła sygnału słabła wraz ze spadkiem tej wartości.

Następnie została wybrana maszyna wektorów nośnych (SVM) z funkcją Gaussa jako funkcją jądrową, ponieważ dawała nam ona najbardziej obiecujące wyniki przy klasyfikacji piętra.

Michał Kowalkowski zajmował się próbą aproksymacji położenia z użyciem SVM, kNN, sieci neuronowych. Jednak i najlepsze wyniki uzyskał podczas pracy z SVMem z jądrem anova.

Wykonano optymalizację parametrów uczenia (w oparciu o wyniki uzyskane w walidacji krzyżowej). Później testowano także różne strategie transformacji danych (przeskalowania pomiarów, poprawki związane z logarytmiczną naturą głośności sygnału, odrzucania pomiarów nieprzekraczających pewnych progów, etc.) Niestety nie miało to istotnego wpływu na wyniki.

3 Praca na danych z MATLABem

Na samym początku zespół Kuby i Dawida zrobił trójwymiarowe mapy w programie MATLAB pokazujące, jak rozkłada się błąd względny na współ-

rzędnych x i y (oddzielnie) dla wszystkich punktów na wydziale dla każdej anteny. Dopasowano kolory (zielony, żółty, czerwony i czarny dla braku sygnału) punktów pomiarów tak, aby było to czytelne i można było dokonać stosownej analizy. Dodatkowo stworzono także mapy błędów w zależności od piętra (oznaczenia podobne jak wcześniej - zielony kolor punktu oznaczał wyznaczenie dobrego piętra, żółty - jedno piętro różnicy, czerwony - różnica większa).

Oprócz błędów, wykonali także trójwymiarowe mapki siły sygnału z anteny. Badanie map błędów i map sił sygnałów pozwoliło jedynie stwierdzić, że błąd w lokalizacji był największy na tych obszarach, gdzie dominuje słaby sygnał.

Warto też wspomnieć o tym, że stworzyli dodatkowo wykres słupkowy, w którym do każdej z anten przyporządkowano sumę sił wszystkich sygnałów odbieranych z tej anteny - to służyło do zorientowania się, czy z danej anteny uzyskano w ogóle jakiś sygnał i jakiej mógł on być jakości. Grupa od uczenia maszynowego początkowo przy klasyfikacji bazowała właśnie na tym, tj. wybierali te anteny, z których był stosunkowo silny sygnał.

3.1 Wycięcie prostopadłościanu z danych

Dawid wyciął z danych prostopadłościan (głównie punkty w okolicy klatki schodowej, usunięte sale z 2. piętra) i stworzył z tego nowy plik z danymi. Miało to służyć sprawdzeniu czy upodobnienie kształtu otoczki punktów, gdzie wykonywano pomiary, do prostopadłościanu, może poprawić osiągane wyniki.

3.2 Wartościowanie anten

Innym zadaniem zespołu była analiza danych celem wyboru anten, których odbierany sygnał rozchodził się monotonicznie w ramach jednej osi. To zdecydowanie ułatwiłoby poprawną lokalizację. Wobec tego należało wprowadzić pewną funkcję wartościującą dla każdej z anteny, aby móc wybrać te, dla których można by z klasyfikacji uzyskać znacznie lepsze wyniki.

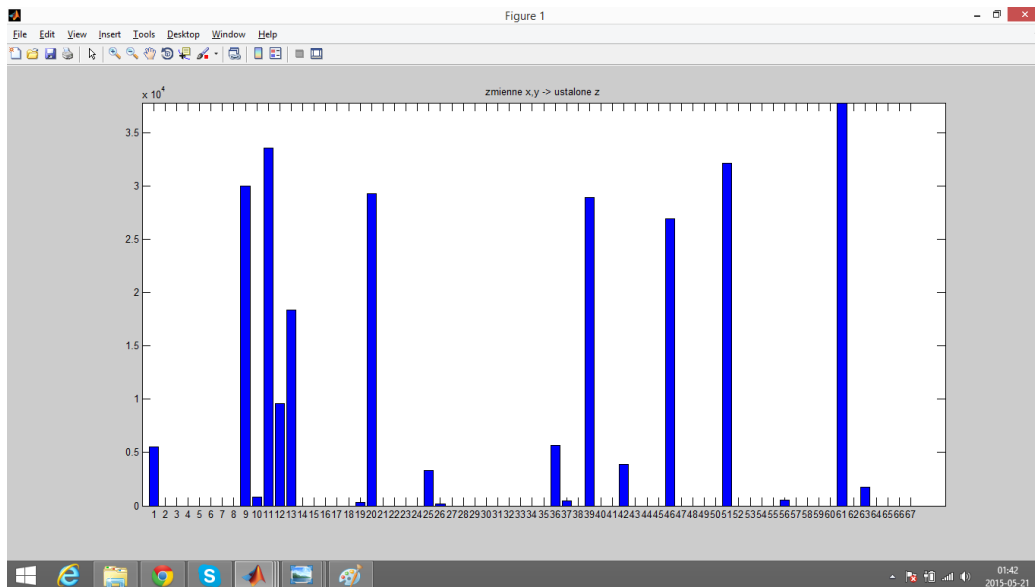
W MATLABie wyznaczenie wartości rankingowej polegało na:

1. wybraniu wszystkich pomiarów w danym punkcie i stworzeniu unikalnego rekordu (te same współrzędne), gdzie mocą anten w tym przypadku była mediana wszystkich mocy sygnałów z każdego punktu - uznaliśmy, że taki model pozwoli na złagodzenie ewentualnych różnic w przypadku wielokrotnych pomiarów
2. następnie należało sprawdzić, jak sygnał zachowuje się w swoistych "tunelach", czyli jak zachowuje się on wzdłuż którejś z osi (OX, OY, OZ) dla ustalonych dwóch współrzędnych niezależnych od przyjętej osi.

Chcieliśmy, by sygnał (w miarę) równomiernie malał albo rósł. Kornel wymyślił sposób, który następnie został zaimplementowany, by liczyć wartość spadków sygnału pomiędzy punktami w sekwencji (na przykład w przypadku pionu spadek sygnału pomiędzy kolejnymi dwoma piętrami) i ilość wzrostów, a następnie bezwzględną różnicę tych wartości. To powodowało, że dla sygnału mniej więcej monotonicznego (czyli dobrego w naszym przypadku) wartość ta była duża, bo był spory wzrost/spadek i bardzo mały spadek/wzrost. Dla sygnału w kształcie "piły zębatej" czy też stałej wartości - w tym przypadku spadki i wzrosty były podobne, dlatego różnica bezwzględna niwelowała to do niemalże zera. Wartością rankingową anten była suma tych różnic po wszystkich punktach (dla pionu po wszystkich x, y).

Co więcej, można zauważyć, że w przypadku braku odbioru sygnału dla dużej ilości punktów wartościowanie będzie niewielkie. Oczywiście zdawaliśmy sobie sprawę, że takie podejście może nie uwzględnić niektórych szczególnych anten, dla których niskie wartościowanie wynika nie ze źle rozchodzącego się sygnału, ale z małej ilości punktów, dla których odebrano sygnał. Wobec tego zastosowaliśmy pomocniczą wartość - maksymalną różnicę w ramach jednego "tunelu", która mogła pomóc w znalezieniu właśnie takich anten i ewentualnym uwzględnieniu ich w procesie uczenia maszynowego.

Mimo wykorzystania tak wyznaczonych rankingów anten do selekcji cech, udało się tylko w niewielkim stopniu polepszyć uzyskiwane wyniki.



4 Szukanie klasyfikatora piętra

Po uzyskaniu początkowo nieobiecujących wyników, zajęliśmy się problemem kategoryzacji anten pod względem różnicy sił sygnałów (zwłaszcza na poszczególnych piętrach). Korzystaliśmy w tym celu z MATLABa generując mapy, które dały nam obraz tego, że w większości wypadków sygnał rozchodzi się na tyle "kuliście" na przestrzeni pięter, że podobne odczyty daje na skrajnych poziomach. To potwierdziły testy na maszynach w Rapidminerze, bo klasyfikator piętra bardzo często mylił właśnie poziomy, które nie sąsiadowały ze sobą, a były równo oddalone od pewnego punktu (np. piętro pierwsze mylone było z piątym).

Później podejmowaliśmy eksperymenty w Rapidminerze na różnych maszynach pod kątem właśnie klasyfikacji piętra i sprawdzaliśmy jak zachowywały się wyniki przy różnych algorytmach wybierania anten (wyrzucania szumiących i zostawiania tych z dużymi różnicami pomiędzy poziomami), ale niestety nie miało to praktycznie żadnego wpływu na wynik. Próbowaliśmy też wykonać ręczną selekcję cech na podstawie przeciętnej jakości sygnału z danej anteny (nie udało się, wyniki były nielepsze niż uzyskane wcześniej).

Jednocześnie próbowaliśmy zmiany podejścia i wytrenowania klasyfikatora nie piętra, a pomieszczenia. To nie dało żadnych praktycznie użytecznych wyników. Warto jednak zwrócić uwagę na fakt, że im bardziej kształt pomieszczenia jest zbliżony do kwadratu (pod warunkiem, że jest tam wystarczająco wiele punktów pomiarowych) tym skuteczniej klasyfikator „przypomina sobie” dane pomieszczenie. Do rozważenia jest opcja sztucznego podziału bu-

dynku na kwadraty i klasyfikowanie kwadratów, żeby wewnątrz nich uszczegółwić wynik.

Należy też zwrócić uwagę na fakt, że dla takiej charakterystyki danych korzystanie z metody kNN prawie automatycznie powoduje "przeuczenie" na zbiór uczący, z uwagi na bardzo dużą liczbę pomiarów w każdym z punktów.

5 Szukanie aproksymatora Y

Nawet SVM nie dał wyników na jakie oczekiwaliśmy - błąd średniokwadratowy dla starych danych - ok. 7, dla nowych ok. 13. Większość pomiarów pochodzi z głównych korytarzy każdego piętra, więc błędy w salach na 2. piętrze były też większe (niedouczenie, z powodu braku porównywalnych danych z podobnych pięter). Dodatkowo braki sygnałów w "losowych" miejscach dawały zbyt duży szum i przez to mieliśmy kłopoty z wyuczeniem maszyn.

Gdy wybrano anteny wyznaczone jako najlepsze przez zespół matlabowy, wyniki się trochę poprawiły (z ok. 8 na 7,2 (błąd dla starych danych)). Odrzucenie 4. i 5. piętra dla nowych danych (bo tam było mniej pomiarów) nie dało żadnych rezultatów - wyniki były porównywalne.

6 Selekcja cech

Opracowana metody rankingowania anten oddzielnie dla każdej osi budynku, z uwzględnieniem monotoniczności jego rozchodzenia się w tej osi, wykorzystana do selekcji cech dla SVM, ostatecznie pozwoliła na uzyskanie najwyższej dotąd skuteczności klasyfikacji piętra na poziomie niecałych 54%. Rozważono dodatkową hipotezę - scenariusz gdzie w "rogach" budynku (wierzchołkach prostopadłościanu) umieszcza się AP WiFi i wykorzystuje sygnał z nich jako dodatkową cechę. Co zaskakujące, nie zaowocowało to istotnie lepszym wynikiem (choć może przy dodatkowym nakładzie pracy można by liczyć na więcej).

7 Zakończenie

Realizacja tego projektu była bardzo pouczająca. Dała nam ona podstawowy przegląd technik uczenia maszynowego (pojęcia takie jak SVM, kro-

swalidacja itp.), pokazała nam potęgę programu RapidMiner i nauczyliśmy się korzystać z jego podstawowych funkcji. Podczas jednych ćwiczeń przeprowadziliśmy eksperyment wspólnego rozpatrywania problemu (nie związanego z naszym projektem) - rozbitków na oceanie. Dzięki tym ćwiczeniom doszliśmy do wniosku, że wspólna analiza wszystkich pomysłów może być bardzo efektywna i przynieść znaczący postęp. Podsumowując, wykonanie projektu było ciekawym doświadczeniem, jednak niestety nie udało nam się go rozwinąć tak, aby nasze rozwiązanie mogło być wykorzystane w praktyce. Być może inne metody (takie jak wykorzystanie więcej niż jednego pomiaru do lokalizacji użytkownika) mogą dać lepsze wyniki, uważamy jednak, że dotarliśmy do granic możliwości zadanego podejścia.