Analiza skorelowanych danych emocjonalnych metodami data mining – streszczenie
praca magisterska, autor inż. Jan Chołoniewski
opiekun: dr inż. Julian Sienkiewicz

Niniejsza praca dyplomowa zawiera wyniki analiz skorelowanych szeregów emocjonalnych metodami statystycznej eksploracji danych *data mining* oraz badanie tych metod pod kątem skuteczności i przydatności do prognozowania przebiegów czasowych aktywności i wartości emocjonalnej na portalu społecznościowym Twitter w czasie wydarzeń sportowych.
Analizowane dane zostały udostępnione przez partnerów projektu CyberEMOTIONS. Składały się one z komentarzy dodanych przez użytkowników serwisu Twitter w promieniu 500 km od Londynu, przed oraz w czasie trwania Igrzysk Olimpijskich w Londynie w 2012 roku i odnoszące się do sportu (było to określane przez etykiety). Każdy komentarz był dodatkowo charakteryzowany przez oszacowaną zawartość pozytywnej oraz negatywnej emocji za pomocą klasyfikatora SentiStrength. Ponadto, do dyspozycji były także daty i godziny zakończenia zawodów finałowych w których Brytyjczycy zdobywali medale. Dane pochodzące z komentarzy agregowano w 15-minutowych oknach czasowych. Określono aktywność (sumę komentarzy w oknie), średnią emocję (średnia arytmetyczna oszacowanych wartości emocjonalnych komentarzy w oknie) i inne dodatkowe wskaźniki.

Tak przygotowanym oknom czasowym były przydzielane klasy w zależności od rozwiązywanego problemu. Zbiory obserwacji i przydzielonych klas służyły do treningu i testowania klasyfikatorów. Wykorzystano w tym celu naiwny klasyfikator Bayesa, liniową i kwadratową analizę dyskryminacyjną, drzewo regresyjne oraz maszyny wektorów podpierających. Oprócz tego przeprowadzono analizę składowych głównych i wykonano klasyfikację, wykorzystując tak uzyskane zmienne.

Działanie klasyfikatorów było analizowane przez zastosowanie ich do rozwiązania trzech problemów klasyfikacyjnych i porównanie parametrów oceniających ich działanie. Za każdym razem testowano wszystkie kombinacje metoda-zestaw zmiennych, przy czym maksymalna liczba zmiennych w zestawie została ustalona na 4 ze względu na ograniczenia czasowe.

Pierwszym zagadnieniem rozwiązywanym przez klasyfikatory było prognozowanie zmiany (spadek/wzrost) szeregu czasowego w kolejnym kroku czasowym. Skuteczność najlepszego uzyskanego klasyfikatora wyniosła 67±2%, co jest dość dobrym wynikiem dla tak złożonego problemu. Z zadaniem zdecydowanie najlepiej poradziły sobie klasyfikatory wykorzystujące metodę maszyn wektorów podpierających z jądrem radialnej funkcji bazowej. Drugim analizowanym problemem było przewidywanie, czy w następnym kroku czasowym aktywność przekroczy określoną wartość progową. Dla tak postawionego problemu, najlepsze klasyfikatory uzyskały 96±1% skuteczności, ale nie były to odosobnione osiągnięcia. Wszystkie metody były w stanie klasyfikować obserwacje z ponad 90% skutecznością dla odpowiednio dobranego zestawu zmiennych. Mimo że wiele metod osiągało tak dobre wyniki, należy wyróżnić maszyny wektorów podpierających z jądrami liniowym, kwadratowym oraz radialnym jako najskuteczniejsze.

Ostatnim badanym problemem było wykrywanie, czy zadane maksimum aktywności mogło zostać wywołane przez istotne sportowe zdarzenie zewnętrzne (tj. zdobycie medalu przez Brytyjczyków). W tym celu, w sygnale aktywności zostały wykryte maksima lokalne, które następnie sklasyfikowano za pomocą informacji o zdobytych medalach do klasy '1' (indukowane przez medal zdobyty przez Brytyjczyków) bądź '1' (występowanie maksimum nie zbiega się w czasie z zakończeniem zawodów, w których Brytyjczycy zdobyli medal). Tak przygotowane dane zostały poddane klasyfikacji, analogicznie jak poprzednio. Wyniki nie wskazują jednoznacznie najlepiej przystosowanego do problemu klasyfikatora. Choć najlepsze uzyskane wartości skuteczności wynoszą 96±1%, to wartość tę należy porównywać z 93%(skuteczność klasyfikatora, który zawsze wskazuje na liczniejszą klasę). W takim wypadku, istotne stają się wartości pomocnicze -- czułość i specyficzność, a także koszty błędnych klasyfikacji, które są ciężkimi do oszacowania parametrami.

Uzyskane wyniki stanowią podstawę do dalszych badań nad wykorzystaniem metod statystycznej eksploracji danych w klasyfikacji emocjonalnych szeregów czasowych.