Analiza danych medycznych z użyciem programu WEKA

1 Cel laboratoriów

Porównanie różnych modeli do drążenia danych wykorzystanych na różnych zbiorach danych medycznych.

Należy wykorzystać program WEKA i wbudowane w nim klasyfikatory.

2 Po co analiza danych medycznych?

Ze względu na wzrastająca liczbę danych o pacjentach powstała potrzeba wydobywania z nich wiedzy co jest podstawowym elementem Medical Decision Support Systems (MDSS's).

Dane medyczne (zawierające informacje o poprzednio zdiagnozowanych chorobach pacjenta) mogą być drążone, by automatycznie wydobyć z nich reguły łączące diagnozę z symptomami. Takie reguły mogą być wykorzystanie do automatycznej klasyfikacji (odnalezienia schorzenia) nowych (niezdiagnozowanych) pacjentów w oparciu o ich symptomy. Ewentualnie mogą być użyte do odnalezienia ukrytej zależności pomiędzy stanem medycznym a faktorami, które na niego wpływają.

3 Dane medyczne

3.1 Specyfika danych medycznych

Dane medyczne zawierają dane o pacjencie (dane ogólne) i zarejestrowane podczas wizyty u lekarza symptomy oraz wyniki testów takich jak pomiar ciśnienia, czy wyniki diagnostyczne (w tym również zdjęcia). Analiza obrazów medycznych jako niezależny temat nie jest przedmiotem niniejszego badania.

Wszystkie dane z baz medycznych muszą być przekonwertowane na dane numeryczne. Bardzo często wartości atrybutów opisujących symptomy są binarne (1— symptom wystąpił, 0 - nie wystąpił).

3.2 Zbiory do wykorzystania w badaniach

Korzystając z `http://archive.ics.uci.edu/ml/` UCI Repository of Machine Learning Databases pobrać następujące zbiory:

1. Heart disease database (ograniczony zbiór z 14 atrybutami)
2. Hepatitis database
3. Dermatology database
4. Wisconsin Diagnostic Breast Cancer (WDBC) (Original)

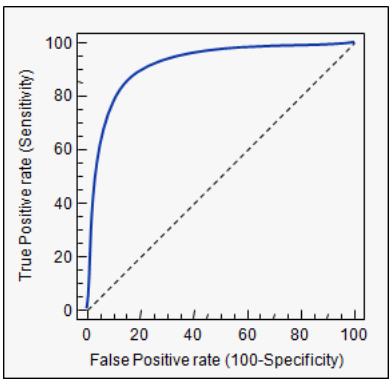
4 Zadania do wykonania

Głównym celem badań jest sprawdzenie jak różne klasyfikatory pracują dla różnych zbiorów medycznych.

4.1 Kryteria porównania klasyfikatorów

Porównując skuteczność algorytmów należy ustalić pewne kryteria porównania. Oto kryteria, które należy wykorzystać w badaniu:

1. Skuteczność predykcji: procent sklasyfikowanych poprawnie elementów. Wady: zbyt optymistyczny. Szczególnie źle wypada, przy nie zbalansowanych danych (90% chorych i 10% zdrowych).
2. Łączny koszt: Średni błąd bezwzględny lub inne błędy średnie.
3. Wrażliwość: w oparciu o błędy: FalsePositive(zdrowi do grupy chorych) and FalseNegative(chorzy do grupy zdrowych).
4. Czułość klasyfikatora wyraża prawdopodobieństwo, że wynik testu będzie pozytywny, gdy pacjent jest chory i wyraża się wzorem $\frac{TP}{TP+FN}$.
5. Specyficzność klasyfikatora wyraża prawdopodobieństwo, że wynik testu będzie negatywny, gdy pacjent nie jest chory $\frac{TN}{TN+FP}$
6. ROC curve (receiver operating characteristic): krzywa czułości (TruePositive rate) w funkcji 100-Specyficzność (FalsePositive rate) dla różnych punktów testowych. Każdy punkt krzywej pokazuje pary czułość/specyficzność odpowiednio do różnych punktów decyzyjnych. ROC w WEKA `http://weka.wikispaces.com/ROC+curves`



Wartość dla Area Under the ROC curve można zinterpretować następująco: jeżeli np. jest równa 0.84, to oznacza, że losowo wybrany pacjent z grupy chorych ma wartość danego testu (dla którego jest rysowana krzywa) większą niż losowo wybranego pacjenta z grupy zdrowych w 84% przypadków. Jeżeli zmienna nie rozróżnia dwóch grup, to wartość pola jest równa 0.5 (leży na przekątnej). Jeżeli rozdzielenie jest idealne (nie ma przecięcia rozkładów), to wartość pola równa 1 (krzywa ROC osiągnie lewy górny róg wykresu).

7. Czas wykonania algorytmu.

4.2 Algorytmy klasyfikacji do zbadania

Należy wykorzystać trzy algorytmy dostępne w pakiecie WEKA:

1. C4.5 (J48)
2. Naiwny klasyfikator Bayesa
3. Wielowarstwowy perceptron.

4.3 Kolejne kroki badania

1. Przyjrzeć się zbiorom danych i ich specyfice. Opisać zbiory w sprawozdaniu.
2. Ustalić jeden model uruchamiania klasyfikatorów (podział na zbiory testowe i uczące). I dobrać do nich parametry. Krótko uzasadnić wybór w sprawozdaniu.
3. Stosując parametry i model wybrany w poprzednim punkcie uruchomić każdy klasyfikator na zbiorach danych i zachować wyniki. Zebrać wszystkie miary opisane w punkcie 4.1. Dodatkowo zachować i przedstawić w sprawozdaniu drzewo decyzyjne z algorytmu C4.5
4. Porównanać klasyfikatory. W czytelny sposób przedstawić wyniki porównania.
5. Skonstruować samodzielnie wnioski z oceny klasyfikatorów.

4.4 Sprawozdanie

Sprawozdanie w formacie i o nazwie *imie_nazwisko.pdf* należy przesłać w terminie do 30-go listopada na adres jkolodziejczy[at]wi.zut.edu.pl. Tytuł maila: Sprawozdanie z ZSIwMET. Opóźnienia będą wpływały na obniżenie punktacji za sprawozdanie.

Wszelkie plagiaty oceniane będą na 0 punktów (niezależnie od autora).

5 Pytania na wejściówkę

1. Jaką specyfikę mają dane medyczne?
2. Jakie elementy może zawierać Medical Decision Support Systems?
3. Jakie stosuje się miary oceny algorytmów klasyfikacji?
4. Jak mierzy się czułość klasyfikatora i co ona wyraża?
5. Jak mierzy się specyficzność klasyfikatora i co ona wyraża?
6. Jaką zależność przedstawia krzywa ROC?
7. Jak interpretować wartości na krzywej ROC?
8. Na czym polega zadanie klasyfikacji?