Poznań 27 października 2009

CASE STUDY 10

Analiza danych medycznych nt. diagnozowania chorób tarczycy

Wprowadzenie

Załóżmy, że współpracujesz ze szpitalem lokalnego uniwersytetu w ramach zastosowań metod informatycznych w medycynie; w szczególności analizy danych. Oddziały tegoż szpitala prowadzą działalność nie tylko w zakresie leczenia pacjentów, szkolenia studentów, ale także badawczą. W działalności badawczej często poszukuje się odpowiedzi na temat nowych procedur diagnostycznych i terapeutycznych. Ma to znaczenie zarówno badawcze jak i szkoleniowe, np. dla studentów czy "młodszych" lekarzy stażystów.

W ostatnim czasie zwrócił się do Ciebie z prośbą o współpracę - konsultację ordynator jednego z oddziałów. Celem była analiza doświadczenia klinicznego w zakresie chorób tarczycy a dokładniej poprawności diagnozowania pewnej odmiany choroby tarczycy.

Poszukiwano modeli charakteryzujących stan pacjentów (chory lub zdrowy) na podstawie danych z wywiadu chorobowego oraz testów diagnostycznych – badania laboratoryjne krwi. Opis problemu medycznego w języku angielskim przedstawiono poniżej.

W ogólności zebrane dane dotyczą 3772 pacjentów opisanych przez 30 atrybutów (zdefiniowanych na skalach zarówno jakościowych jak i ilościowych). W przypadku części obserwacji/pacjentów część atrybutów zawiera niezdefiniowane wartości.

Definicja dziedzin atrybutów przedstawiona jest poniżej

age:	continuous.
sex:	M, F.
on thyroxine:	f, t.
query on thyroxine:	f, t.
on antithyroid medication:	f, t.
sick:	f, t.
pregnant:	f, t.
thyroid surgery:	f, t.
I131 treatment:	f, t.
query hypothyroid:	f, t.
query hyperthyroid:	f, t.
lithium:	f, t.
goitre:	f, t.
tumor:	f, t.
hypopituitary:	f, t.
psych:	f, t.
TSH measured:	f, t.
TSH:	continuous.
T3 measured:	f, t.

T3:	continuous.
TT4 measured:	f, t.
TT4:	continuous.
T4U measured:	f, t.
T4U:	continuous.
FTI measured:	f, t.
FTI:	continuous.
TBG measured:	f, t.
TBG:	continuous.
referral source:	WEST, STMW, SVHC, SVI, SVHD, other.
classes	sick, negative.

Ogólna informacja o zbiorze danych:

Liczba obserwacji: 3772
Łączna liczba atrybutów: 30
W tym liczbowych: 7 (Int 1 / Real 6)
oraz jakościowych: 23
Liczba tzw. missing values of attribute: 6064 / 5.4%
Rozkład liczności klas: negative – zdrowy 3541 oraz sick (diagnoza choroby) 231 pacjentów.

Zwróć uwagę, że dane są silnie niezrównoważone. Z medycznego punktu widzenia rozpoznawanie choroby jest ważniejsze niż ew. pomyłki w stosunku do zdrowych osób. Dlatego w swojej analizie powinieneś zapewniać jak najlepsze postępowanie do stosunku do rozpoznawania klasy choroby (zainteresuj się metodami analizy tzw. z ang. imbalanced data).

Powinieneś pamiętać, iż nie masz wpływu na rozmiar dostępnych danych, nie możesz oczekiwać dostarczenia dodatkowych opisów pacjentów; zostało to wykonane przed Twoim udziałem w studium badawczym; Tzn. nie będziesz miał dodatkowych obserwacji lub nie jest możliwe wprowadzenie dodatkowych atrybutów.

Natomiast, jeśli potrafisz ocenić jakość otrzymanych danych możesz dokonywać przeskalowań lub prze-definiowań atrybutów (np. tworzyć nowe atrybuty w oparciu o pomierzone), jeśli ich końcowa postać jest akceptowalna dla potencjalnego użytkownika (czytaj ma potencjalnie dogodną interpretacje medyczną).

Inne uwagi metodyczne:

Opłaca się badać jakość dostarczonych danych (były one zbierane przez osoby, które nie znają podstaw Twoich metod eksploracji danych);

- Interesujące jest badanie wzajemnych współzależności tkwiących w danych;
- Należy dokonać próby oceny ważności poszczególnych atrybutów. I jeśli wyniki tej analizy będą jednoznaczne może warto dokonać selekcji najważniejszych z nich – ale zachowaj tutaj dużą ostrożność, gdyż zostały one już poprzednio wybrane przez ekspertów,
- Warto stosować więcej niż jedną metodę eksploracji danych (ukierunkowanych na różne formy wiedzy i różne ich reprezentacje)

- Uwaga z powodu silnego eksploatowania na zajęciach dotychczas technik budowy klasyfikatorów zwracam uwagę, że automatyczna klasyfikacja pacjentów nie jest zawsze akceptowana przez wielu lekarzy; może być używana jednak jako miara pewności czy wiarygodności wyników;
- Interesujące może być poszukiwanie reprezentacji zależności pomiędzy wartościami wybranych atrybutów warunkowych a decyzyjnym w postaci symbolicznej, np. drzew lub reguł decyzyjnych, a następnie próba zbudowania modelu charakterystycznego pacjenta z danej klas.
- Jeśli chcesz rozważać także budowę klasyfikatorów, Lekarze nie skupiają się na globalnej trafności klasyfikacji lecz ważniejsza jest dla nich trafność w poszczególnych klasach, w szczególności osób chorych = klasa sick (analiza "confusion matrix" jest bardzo pożadana). Należy wtedy przyjąć oczekiwanie lekarzy mówiące, że skuteczność rozpoznawania pacjentów z tej klasy powinna być nie niższa niż **92%** ! Natomiast globalne trafność powinna być równocześnie powyżej 95%.
- Metodycznie należy patrzeć na ten problem, jako uczenie się klasyfikacji z silnie niezrównoważonych klas decyzyjnych, co może oznaczać poszukiwanie niestandardowych metod.

Inne zagadnienia przekaże prowadzący – konsultant.