Poznań 7 listopada 2008

CASE STUDY 1

Analiza danych medycznych nt. leczenia pacjentów cierpiących na chorobę wrzodową dwunastnicy.

(***Copyright*** Jerzy Stefanowski -, Instytut Informatyki PP;
Zastrzeżenia dotyczą opisu problemu i ograniczonej dostępności do danych)

Cel :

„Case study" powinien prowadzić do odkrycia użytecznych i potencjalnie interesujących regularności z rzeczywistych danych. Należy także dokonaniu interpretacji i oceny znalezionych regularności. Możliwie jest interpretowanie znalezionych regularności jako form reprezentacji wiedzy odkrytych w bazie danych. Problem dotyczy analizy medycznej bazy danych. Z metodologicznego punktu widzenia sugeruje się wykorzystywanie poznanych metod analizy danych zarówno statystycznych jak i wywodzących się ze sztucznej inteligencji metod nadzorowanego uczenia maszynowego.

Podsumowaniem powinien być raport zawierający zestawienie najbardziej interesujących zależności, regularności czy zaskakujących anomalii oraz komentarz zawierający ich interpretację.

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 skuteczności nowych leków, stosowania określonych 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 oddziału chirurgicznego. Problem dotyczy analizy stanu zdrowia pacjentów cierpiących na chorobą wrzodową dwunastnicy, leczonych metodą wysoce wybiórczej wagotomii (ang. Highly Selective Vagotomy – w skrócie *HSV*). Opis medyczny załączony na dodatkowych materiałach – patrz Załączniki nr. 1.

Dane dotyczą pacjentów przyjmowanych do szpitala i skierowanych do zabiegu chirurgicznego wysoce wybiórczej wagotomii. Pacjenci są opisani 15 atrybutami charakteryzującymi wyniki wywiadu chorobowego, badań przedmiotowych oraz analizy wydzielin żołądkowych (różne parametry wydzielania soku żołądkowego, także po pobudzeniu podaniem tzw. testu histaminowego) przeprowadzonych przed zabiegiem. Są to następujące atrybuty:

*A*1: - Płeć

*A*2: - Wiek [lata]

*A*3: - Okres trwania choroby [lata]

*A*4: - Bolesność uciskowa w nadbrzuszu [T- tak/ N - nie]

*A*5: - Ból w nadbrzuszu po zjedzeniu posiłku występujący w czasie 1 godziny [N – brak, S – słaby, L – silny].

*A*6: - Występowanie u pacjenta odczucia tzw. „zgagi" czy „odbijania" [- tak/ N - nie]

*A*7: - Obecność „niszy" na zdjęciu RTG {T - tak/ N - nie]

*A*8: - Komplikacje wrzodu [N – brak, A – ostry krwotok, M - wielokrotne krwawienie, PE - perforacja, PY - zwężenie odźwiernika

*A*9:- Koncentracja HCL [mmol HCL/100ml]

A10 - Objętość wydzielania soku żołądkowego na godzinę [ml/godz]

*A*11 – Objętość wydzielania soku żołądkowego zalegająca [ml]

*A*12 – Tzw. BAO – ang. Basic Acid Output [mmol HCL/100ml]

Podobne parametry jak *A*9, *A*10, *A*12 ale po wykonaniu testu histaminowego

*A*13:- Koncentracja HCL [mmol HCL/100ml]

*A*14 – Objętość wydzielania soku żołądkowego na godzinę [ml]

*A*15 – Tzw. MAO – ang. Maximal Acid Output [mmol HCL/100ml]

Długoterminowy rezultat zabiegu jest oceniany według, tzw. skali Visick'a, w czterowartościowej skali – wyraża ona skuteczność stosowanego leczenia:

1. Znakomity wynik leczenia.
2. Bardzo dobry.
3. Satysfakcjonujący.
4. Niesatysfakcjonujący.

Można przyjąć, że jest to atrybut decyzyjny, który definiuje klasyfikację pacjentów w rozważanej tablicy decyzyjnej. Klasy w sensie liczności są niezrównoważone (silna przewaga grupy wyleczonych pacjentów). Atrybuty są różnego typu (nominalnego, porządkowego jak i liczbowego). Dostępne są także normy medyczne interpretacji wyników testów, które mogą być podstawą eksperckiej dyskretyzacji atrybutów liczbowych – Patrz Załącznik nr. 1.

W obecnej chwili ordynator X i jego współpracownicy zebrali dane o 186 pacjentach, patrz załączony zbiór danych.

Problem, jaki sobie stawiają lekarze, dotyczy poszukiwania istotnych zależności, regularności pomiędzy atrybutami, a także ustalania hierarchii ważności atrybutów opisujących pacjentów. Prowadzi to do oceny znaczenia atrybutów i ich podzbiorów dla przybliżenia klasyfikacji pacjentów; poszukiwania tzw. modeli pacjentów charakterystycznych dla klas decyzyjnych; a także określenie wskazań do przeprowadzania zabiegu chirurgicznego wysoce wybiórczej wagotomii (*HSV*) dla pacjentów cierpiących na chorobę wrzodową dwunastnicy.

Powyższe badanie ma przede wszystkim doprowadzić do powstania raportu naukowego. Mogą być wykorzystane także do celów szkolenia studentów. Konieczna jest, więc w miarę formalna analiza i uzasadnianie wniosków.

Możesz założyć, ze aktualne doświadczenie zespołu medycznego obejmuje przede wszystkim doświadczenie w zakresie statystycznej analizy danych a także podstawową wiedzą w zakresie metod wspomagania decyzji, np. maszynowego uczenia się oraz „data mining". Lekarze mogli uczestniczyć w specjalnych szkoleniach, gdzie ogólnie omawiano te zagadnienia.

W załącznikach otrzymujesz:

1. Opis problemu medycznego wraz z informacją o możliwych normach interpretacji wartości atrybutów liczbowych.

2. Plik z danymi (format tekstowy isf lub plik Excel'a).

3. Krótki tekst z pewnego podręcznika na temat eksploracji danych medycznych i ograniczeń tego typu działań.

Uwagi metodyczne:

Powinieneś pamiętać, że nie masz wpływu na rozmiar dostępnych danych, nie możesz oczekiwać dostarczenia dodatkowych opisów pacjentów; zostało to wykonane przed Twoim udziałem w studium badawczym (dane oryginalnie zebrano w innym celu, np. dokumentacji historii chorób, chorób dopiero później uznano, że mogą podlegać analizie – ang. secondary data analysis); Tzn. nie będziesz mógł żądać dodatkowych obserwacji lub wprowadzenia dodatkowych atrybutów. Natomiast, jeśli potrafisz ocenić jakość otrzymanych danych możesz dokonywać przeskalowań lub przedefiniować atrybutów (np. tworzyć nowe w oparciu o pomierzone), jeśli ich końcowa postać jest akceptowalna dla potencjalnego użytkownika (czyli ma potencjalnie dogodną interpretacje medyczną).

Inne uwagi metodyczne:

Opłaca się badać jakość dostarczonych danych (mogą być zbierane przez osoby, które nie znają własności Twoich metod);

- Interesujące jest badanie wzajemnych współzależności tkwiących w danych;
- Nie wszystkie atrybuty mogą być związane z klasyfikacją pacjentów (niektóre są wręcz nadmiarowe lub niepotrzebnie zarejestrowane) – warto stosować metody wstępnego przetwarzania danych.
- Warto stosować więcej niż jedną metodę eksploracji danych (ukierunkowanych na różne formy wiedzy i różne ich reprezentacje).
- Interesujące może być poszukiwanie reprezentacji zależności pomiędzy wartościami wybranych atrybutów warunkowych a decyzyjnym w postaci symbolicznej, np. drzew lub reguł decyzyjnych, albo funkcji dyskryminacyjnych (ang. Fisherian discriminant analysis).

- Uwaga!; z powodu silnego eksploatowania na zajęciach dotychczas technik budowy klasyfikatorów, zwracam uwagę, że automatyczna klasyfikacja pacjentów nie jest na ogół akceptowana przez wielu lekarzy; może być używana jednak jako miara pewności czy wiarygodności wyników.
- Jeśli chcesz rozważać także budowę klasyfikatorów, to pamiętaj, że lekarze nie skupiają się na globalnej trafności klasyfikacji, lecz ważniejsza jest dla nich trafność w poszczególnych klasach, w szczególności dla osób w gorszym stanie (analiza "confusion matrix" jest bardzo pożądana).
- W medycznych danych występują często niezrównoważone klasy decyzyjne oraz obserwacje samotnicze - należy je właściwie uwzględniać.

---------- **Załącznik nr 1**

Interpretacja norm medycznych mogąca być podstawą eksperckiej dyskretyzacji niektórych z atrybutów

Na podstawie pracy dotyczącej leczenia choroby wrzodowej dwunastnicy:

*A*2 [lata]: ≤ 35, >35

*A*3 [lata]: ≤ 0.5, (0.5,3], >3

*A*9 [mmol HCL/100ml]: ≤ 2, (2,4], >4

*A*10 [ml]: ≤ 70, (70,150], >150

*A*11 [mmol/100ml]: ≤ 50, (50,150], >150

A12 [mmol HCL/100ml] ≤ 2, (2,3], >3

*A*13 [mmol HCL/100ml]: ≤ 10, (10,15], >15

*A*14 [ml]: ≤ 100, (100,250], >250

*A*15 [mmol/100ml]: ≤ 15, (15,25], >25

---------- **Załącznik nr 2**

Analiza danych medycznych (wyciąg z artykułu przeglądowego Z.Pawlak, K.Słowiński i J.Stefanowski)

Przedstawiane problemy dotyczą eksploracji danych dla wspomagania decyzji w zakresie diagnozy i doboru terapii czy sposobu postępowania medycznego. W ogólności podejmowanie decyzji diagnostycznych w medycynie jest skomplikowanym procesem, który obejmuje zbieranie różnych danych i ich interpretacje przez lekarza. Dane mogą pochodzić z różnych źródeł, np. wywiadu chorobowego, badań przeprowadzonych przez lekarza czy wykonania szeregu testów laboratoryjnych lub procedur diagnostycznych (obrazowych, sygnałowych, itp.). Analiza tych danych, jak i doświadczenie kliniczne lekarza powinno prowadzić do postawienia diagnozy (rozpoznania wstępnego i później końcowego) oraz do określenia właściwej terapii. Zadanie to jest trudne, jeśli aktualny stan wiedzy na temat danej choroby nie jest jeszcze dostatecznie rozwinięty lub nie ma jednomyślności pomiędzy specjalistami. Ponadto, postęp technologiczny w zakresie urządzeń diagnostycznych zwiększył znacząco rozmiar dostępnych danych pomiarowych. W obu przypadkach zebrane dane mogą mieć, więc różne znaczenie. Podlegają one dalszej analizie w celu odnalezienia i wybrania najważniejszych elementów dla interpretacji medycznej. Typowe zadania z medycznego punktu widzenia to:

1. Identyfikacja najważniejszych atrybutów dla klasyfikacji pacjentów,
2. Odkrywanie zależności pomiędzy wartościami atrybutów a przydziałem pacjenta do określonych klas.