Case study 24 z przedmiotu „Zaawnsowana Eksploracja Danych" – TPD (listopad 2010)

Analiza preferencji wyborczych

(***Copyright*** **Jerzy Stefanowski** - Instytut Informatyki Politechnika Poznańska; 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. Dość ważną częścią studium przypadku jest dokonanie oceny ważności atrybutów i ustalenie które czynniki najsilniej wpływają na decyzje wyborcze. W ogólności chcemy stworzyć model typowego wyborcy danej partii. Możliwie jest interpretowanie znalezionych regularności jako form reprezentacji wiedzy odkrytych w bazie danych. W zasadzie nie chodzi o budowę automatycznego klasyfikatora, lecz o stworzenie charakterystycznego opisu preferencji wyborczych. Problem dotyczy danych ankietowych z badania wyborców amerykańskich.
Z metodologicznego punktu widzenia sugeruje się wykorzystywanie poznanych metod eksploracji danych i odkrywania wiedzy - zarówno statystycznych jak i wywodzących się ze sztucznej inteligencji.

Podsumowaniem analizy powinien być krótki raport zawierający listę najbardziej interesujących regularności oraz komentarz lub ich interpretację - raport ten powinien być tworzony na bieżąco podczas zajęć.

Dane i wymagania dla eksploracji danych:

Dysponujemy badaniami ankietowymi wyborców w USA głosujących w wyborach prezydenckich na kandydatów partii republikańskiej i demokratycznej. Ankiety zawierają informacje o odpowiedziach na 30 pytań (atrybuty w dostarczonym pliku). Wartości odpowiedzi są punktowane wg. informacji o skalach dostarczonych dodatkowo od prowadzącego (najczęściej jest to ocena odległości ideologicznej od kandydata w danym problemie, np. ochronie zdrowia). Powyższe atrybuty mają w większości charakter porządkowy. Możliwe jest wystąpienie braku pewnej odpowiedzi (kodowane jako „?"). Dane mogą także zawierać pewne błędy – osoba przygotowująca plik nie jest zbyt wprawna w obsłudze komputera.

Cele eksploracji danych obejmują:

(**a**) ocena jakości dostarczonych danych i ew. oczyszczanie.

(**b**) ocena przydatności pojedynczych atrybutów dla oceny, na kogo głosuje wyborca – warto stworzyć pewien rodzaj rankingu ważności atrybutów (i to z wykorzystaniem różnych metod),

(**c**) ewentualna selekcja zbioru atrybutów do podzbioru najważniejszych zapewniających satysfakcjonującą ocenę wyborcy,

(**d**) poszukiwanie zależności między wartościami wybranych atrybutów a decyzją wyborczą (republikanin vs. demokrata),

(**e**) konstrukcja tzw. profilu wyborcy.

W przypadku rozważanego case study dostarczono Tobie gotowy zbiór danych i poproszono o jego jak najintensywniejszą eksplorację. Nie jesteś w stanie zażądać już dokonania dodatkowych badań ankietowych– musisz starać się odkryć jak najwięcej interesujących elementów w danych, które otrzymałeś.

Twoje wnioski powinny być interpretowane w kategoriach przydatności odkrytej wiedzy z danych (czyli znalezionych regularności, ważności atrybutów itp.) dla oceny czym się różnią wyborcy popierający kandydata republikanów od wyborców kandydata demokratów.

Ponadto można badać współzależności pomiędzy samymi atrybutami, aby wykryć wzajemne uwarunkowania. Uwaga stworzenie klasyfikatora nie jest celem tego studium. Lecz jeśli będziesz to wykonywał pamiętaj, że ew. klasyfikator ma być tylko pomocniczy dla oceny modelu wyborcy z danej partii. Główny cel tego badania to perspektywa opisowa eksploracji danych a nie predykcji decyzji klasyfikacyjnej.

Dane są dostępne jako plik ASCII o formacie tekstowym w zapisie jeden wiersz zawierający opis jednego pojazdu za pomocą powyższych symptomów i atrybutów decyzyjnych. W pierwszej części pliku dostępny jest nagłówek zawierający definicje dziedzin atrybutów. Plik może zawierać błędy wykonane przez osobę wprowadzającą dane.

Dodatkowe uwagi metodyczne:

Powinieneś pamiętać, iż nie masz wpływu na rozmiar dostępnych danych, nie możesz oczekiwać dostarczenia dodatkowych opisów przypadków; wszystko zostało to wykonane przed Twoim udziałem w studium badawczym - nie możesz żądać dodatkowych czynności pozyskiwania informacji. Ewentualnie dowiadując się, o jakie konkretnie wybory chodziło możesz samodzielnie poszukać właściwych materiałów uzupełniających.

Jest to typowa eksploracja dostępnych danych ukierunkowana na stworzenie opisu danych.

Konieczne jest badanie jakości dostarczonych danych (mogą być zbierane przez osoby, które nie znają własności Twoich metod); Ponadto podczas przygotowywania danych mogły wystąpić pomyłki wprowadzania pomiarów

Metodycznie potraktuj problem jako odpowiednie zadania tzw. uczenia nadzorowanego (dany jest opis przykładów za pomocą atrybutów jak i klasyfikacja – lecz możesz też w ramach opisu atrybutami rozważać zdanie nienadzorowane) ukierunkowane na tzw. symboliczne reprezentacje wiedzy. Rozważaj także właściwe metody statystyczne.

Inne uwagi metodyczne:

- Interesujące jest badanie wzajemnych współzależności tkwiących w danych;
- Analizuj każdy atrybut warunkowy oddzielnie a później ich właściwe podzbiory.
- Warto stosować więcej niż jedną metodę eksploracji danych (ukierunkowanych na różne formy wiedzy i różne ich reprezentacje)
- Uwaga z powodu silnego eksploatowania na dotychczasowych zajęciach technik budowy klasyfikatorów, zwracam uwagę, że automatyczna klasyfikacja nie jest celem tego studium.

Dostępne oprogramowanie:

Oprogramowanie Statystyczne Statistica, R, itp.

Środowiska jak WEKA, RapidMiner, i inne.

System indukcji drzew decyzyjnych C4.5; System indukcji reguł decyzyjnych CN2; System oparty na teorii zbiorów przybliżonych ROSE

Inne wg. uznania

Proponowany przebieg zajęć:

1. W pierwszej części "Case Study" prowadzący omawia problemy eksploracji danych w diagnostyce technicznej; następnie krótko charakteryzuje poniższy problem.
2. Uczestnicy zapoznają się z niniejszym tekstem i danymi, starając się określić cel i zakres swojej analizy oraz zidentyfikować podstawowe właściwości danych; Ponadto starają się przygotować przed zajęciami plan swoich zamierzeń. Powinien on obejmować zakładany cel analizy; listę problemów diagnostycznych, które zamierza się rozwiązać i powiązanych z tym interesujących typów regularności, których zamierza się poszukiwać w danych. Należy także określić podstawowe własności danych i listę metod, które się zamierza użyć.Uczestnicy są podzieleni na zespoły i wewnatrz zespołów dyskutują na temat problemu oraz przygotowują propozycje rozwiązania. Zespoły uczestników powinny **realizować samodzielnie analizę i starają się na bieżąco prowadzić raport z wykonywanych czynności i uzyskanych wyników**.
3. Zespoły w dyskusji (**druga część spotkania**) prezentują propozycje rozwiązania problemu, jak i także omawiają wyniki samodzielnej analizy – zalecanie wykorzystanie tworzone **raportu - sprawozdania** z dotychczasowej analizy i otrzymanych rezultatów.Na końcu zajęć prowadzący omawia się wspólnie otrzymane wyniki, prowadzący prezentuje także inne znane wyniki dla tego przypadku. Prowadzi się dyskusję na ich temat i podsumowuje całość zajęć.
4. Zespoły dostarczają sprawozdzanie końcowe. Sprawozdanie, aktywność na zajęciach oraz obecność na tych zajęciach jest podstawą otrzymania oceny za przebieg tego case study.