Case study 12 z przedmiotu „Zaawansowana Eksploracja Danych” – V rok TPD (listopad 2008)

Analiza diagnostycznej bazy 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 stanu technicznego autobusów używanych przez jedno z przedsiębiorstw w Polsce.
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.

Wprowadzenie :

Pewne przedsiębiorstwo komunikacyjno-transportowe wykorzystuje jeden podstawowy typ autobusów. Autobusy te podlegają regularnym przeglądom stanu technicznego. Zbiera się również dane o ich bieżącej eksploatacji. Na stanowisku diagnostycznym bada się przede wszystkim parametry układu napędowego, takie jak aktualna maksymalna moc silnika, moment obrotowy, osiągana prędkość maksymalna, ciśnienie sprężania w tłokach, zawartość elementów smołowatych w spalinach i inne. Dane eksploatacyjne dotyczą głównie zużycia paliwa i oleju w różnych okresach roku.

Na podstawie uzyskanych wyników testów diagnostycznych i informacji eksploatacyjnych podejmuje się decyzje operacyjne, co do dalszego użytkowania pojazdów. Decyzje te mogą być różne, np. pojazd w dobrym stanie technicznych, wymagający szczególnej obserwacji, ale warunkowo dalej eksploatowany, wycofany z eksploatacji i przeznaczony do remontu.

Część danych posiada istniejące wskazówki producenta, tj. tzw. wartości graniczne symptomów stanu technicznego – mają one na celu identyfikacje przedziałów „normalnych” wartości stanu technicznego oraz przedziałów świadczących o pogorszeniu się stanu technicznego.

Należy zwrócić uwagę, że symptomy stanu technicznego mogą mieć różną przydatność dla całościowej oceny stanu pojazdów. Niektóre z nich niosą wartościową informację. Inne mają mniejszy wpływ dla ostatecznej oceny. W wielu przypadkach rozważa się większą liczbę symptomów niż jest to konieczne.

Celem badań diagnostycznych jest:

(**a**) ocena przydatności diagnostycznych pojedynczych symptomów oraz wartości granicznych,

(**b**) redukcja zbioru symptomów do podzbioru najważniejszych zapewniających satysfakcjonującą ocenę stanu technicznego,

(**c**) poszukiwanie zależności między wartościami wybranych symptomów a globalną oceną stanu technicznego,

(**d**) konstrukcja tzw. klasyfikatora stanu technicznego, czyli wskazań dla podejmowania końcowej decyzji diagnostycznej na podstawie bieżących obserwacji.

Ponadto można badać współzależności pomiędzy samymi symptomami stany technicznego, wybór najważniejszych symptomów, czy dokonywać oceny klasyfikacyjnej tworzonych klasyfikatorów.

W przypadku rozważanego case study dostarczono Tobie gotowy zbiór danych i poproszono o przebadanie w celu odpowiedzenia na powyższe pytania i problemy związane z zagadnieniami diagnozowania stanu technicznego pojazdów.

Twoje wnioski powinny być interpretowane w kategoriach przydatności odkrytej wiedzy z danych (czyli znalezionych regularności, klasyfikatorów itp.) dla bieżącej oceny diagnostycznej podczas badań stanu technicznego pojazdów i podejmowania decyzji o ich dalszej eksploatacji na podstawie wartości wybranych symptomów stanów technicznego.

Dane dotyczą autobusów tego samego typu eksploatowanych w podobnych warunkach. Zawierają one wartości następujących parametrów, będących symptomami stanu technicznego:

s1 –prędkość maksymalna [km/h],

s2 –ciśnienie sprężania [Mpa],

s3 – zawartość elementów smołowatych w spalinach wylotowych [%],

s4 – moment obrotowy silnika [Nm],

s5 – letnie zużycie paliwa [l/100km],

s6 – zimowe zużycie paliwa [l/100km],

s7 – zużycie oleju [l/1000km],

s8 –aktualna moc silnika [KM].

Innych parametrów nie umieszczono w posiadanym zbiorze danych.

Wszystkie pomiary są zdefiniowane na skalach numerycznych. W ogólności jest też możliwe ich interpretowanie w kategoriach jakościowych w oparciu o wartości graniczne – patrz dodatek.

Symptomy takie jak s5, s6, s7 dotyczą ogólnej oceny eksploatacji autobusów a pozostałe oceniają stan silnika. Ponadto dostępne są dwie propozycje atrybutów decyzyjnych wyrażających dwie możliwe decyzje diagnostyczne:

1. Pierwszy atrybut klasyfikuje pojazdy na dwie klasy sprawne [kod „1”] i niesprawne [kod „2”].
2. Drugi atrybut decyzyjny klasyfikuje pojazdy jako sprawne, wymagające drobnych napraw i obserwacji oraz przeznaczone do remontu (kodowane odpowiednio 1,2,3 – im wieksza wartość tym mniej sprawny pojazd).

Dane są dostępne jako plik ASCII o formacie ISF, w zapisie jeden wiersz zawierający opis jednego pojazdu za pomocą powyższych symptomów i atrybutów decyzyjnych. Plik zostanie udostępnione przez prowadzącego.

Tymczasowo dostępny jest poprzez **stronę WWW o adresie**:

http://www-idss.cs.put.poznan.pl/~stefan/aed/autobusy.isf

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.

Jest to typowa eksploracja dostępnych danych.

Jeśli potrafisz ocenić jakość otrzymanych danych możesz dokonywać przeskalowań lub przedefiniowań atrybutów (w oparciu o określone metody).

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

- warto tutaj skontrolować, czy nie występują pomiary o wartościach mocno odległych od innych typowych wartości (tak zwane obserwacje odstające),
- sprawdź także, czy nie ma błednych lub nieznanych wartości niektórych atrybutów = w wyniku badań diagnostycznych starano sie dokonać wszystkich pomiarów.

Metodycznie potraktuj problem jako odpowiednie zadania tzw. uczenia nadzorowanego (dany jest opis przykładów za pomocą atrybutów jak i klasyfikacja). Rozważaj także właściwe metody statystyczne

Inne uwagi metodyczne:

- Interesujące jest badanie wzajemnych współzależności tkwiących w danych;
- Analizuj każdy atrybut decyzyjny 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 jest tylko jedyną z miar oceny; Jeśli chcesz rozważać budowę klasyfikatorów to pamiętaj także iż użytkownicy nie skupiają się wyłącznie na globalnej trafności klasyfikacji lecz ważniejsza jest dla nich trafność w poszczególnych klasach, w szczególności pojazdów w gorszym stanie (analiza „confusion matrix" jest bardzo pożądana).

Dostępne oprogramowanie:

Oprogramowanie Statystyczne

Pakiet EXCEL - moduły analizy danych

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. Zespóły dostarczają sprawozdzanie końcowe. Sprawozdanie, aktywność na zajęciach oraz obecność na tych zajęciach jest podstawą otrzymania oceny za przebieg tego case study.

Dodatek A:

Symptomy mające w oryginale wartości liczbowe zostały przetransformowane na wartości dyskretne porządkowe w oparciu o wartości graniczne; Poniżej propozycja pochodząca od eksperta diagnostyki samochodowej:

s1 : *(-∞, 74>, (74, 79>, (79, 85>, (85, ∞)*

s2 : *(-∞, 2.2>, (2.2, 2.4>, (2.4, ∞)*

s3 : *(-∞, 59>, (59, ∞)*

s4 : *(-∞, 44.1>, (44.1, ∞)*

s5 : *(-∞, 22>, (22, ∞)*

s6 : *(-∞, 25.2>, (25.2, ∞)*

s7 : *(-∞, 1.2>, (1.2, ∞)*

s8 : *(-∞, 119>, (119, ∞)*

Jeśli to potrzebne możesz rozważać powyższą propozycję, ale podstawową analizę proszę wykonać dla danych zdefiniowanych na skalach numerycznych.

Dodatek B:

Problemy diagnostyki technicznej – podstawowe informacje

Przedstawimy kilka uwag na temat celów diagnostyki technicznej i zadań, w których rozwiązaniu mogą być przydatne metody odkrywania wiedzy oraz eksploracji danych. Efektywna eksploatacja obiektów mechanicznych czy maszyn w procesach przemysłowych wymaga wiarygodnej informacji o ich stanie technicznym. Informacja ta jest często rozszerzona o predykcję zmiany ich stanu technicznego. Ma to szczególne znaczenie w przypadku maszyn o krytycznym znaczeniu w procesie, wymagających specjalnych skomputeryzowanych systemów nadzoru (monitorowania) – np. turbiny w elektrowniach. W przypadku maszyn o mniejszym znaczeniu dokonuje się okresowych pomiarów za pomocą urządzeń przenośnych. Ocena stanu technicznego wykonywana jest na podstawie tzw. symptomów, czyli wielkości, które zmieniają się wraz z pogarszaniem się stanu technicznego maszyny. Przykładami symptomów są poziom drgań reprezentatywnych punktów obiektu, poziom hałasu w otoczeniu pracującej maszyny, temperatura, ciśnienie czy wzajemna pozycja części składowych maszyny. Z diagnostycznego punktu widzenia poszczególne symptomy mają różną przydatność dla konkretnego problemu. Dla pojedynczych symptomów mierzonych na skalach liczbowych definiuje się czasami tzw. wartości graniczne. Dzielą one dziedzinę symptomu na pewne podprzedziały, które mogą być interpretowane w kategoriach normalnych warunków pracy i ich stopniowego pogarszania się. Są one definiowane na podstawie zaleceń producentów i wskazań literaturowych, lecz praktyczne badania nie zawsze potwierdzają ich użyteczność. Ponadto w postępowaniu diagnostycznym mamy zwykle dostęp do dużych zbiorów danych. Dane gromadzone podczas obserwacji obiektów diagnozowania mogą być także niekompletne, sprzeczne, niedokładne czy obarczone niepewnością. Także sama wiedza diagnostyczna ma często charakter przybliżony. Dyskusja tych zagadnień przedstawiona jest w [Moczulski 97]. Z punktu widzenia zastosowań rozważanych w tym rozdziale istotnymi zadaniami badań diagnostycznych są :

1. ocena zdolności diagnostycznej poszczególnych symptomów,
2. ocena różnych metod definiowania wartości granicznych dla tych symptomów,
3. selekcja podzbioru symptomów zapewniających satysfakcjonującą ocenę stanu technicznego,
4. stworzenie klasyfikatora stanu technicznego.

Moczulski W., Metody pozyskiwania wiedzy dla potrzeb diagnostyki maszyn. Zeszyty Naukowe PolitechnikiŚląskiej. Monografie, Mechanika, z. 130, Gliwice 1997.