Analiza Wielowymiarowa

Analiza dyskryminacyjna i analiza skupień

Maciej Nasiński, Paweł Strawiński

Zajęcia 6
10 listopada 2022

Plan zajęć

1. Analiza Dyskryminacyjna
   - O analizie
   - Przykład
2. Analiza Skupień
   - O analizie
   - Przykład

Analiza dyskryminacyjna

- Analiza dyskryminacyjna jest stosowana do rozstrzygania, które zmienne wyróżniają (dyskryminują) dwie lub większą liczbę naturalnie wyłaniających się grup
- Szuka reguły przyporządkowania wielowymiarowych obiektów do jednej z wielu klas przy możliwie minimalnych błędach klasyfikacji
- Główna idea to rozstrzyganie, czy grupy różnią się ze względu na średnią wartość pewnej cechy (zmiennej), a następnie wykorzystanie tej zmiennej do przewidywania przynależności do grupy

Cel analizy

- Celem opisowej analizy dyskryminacyjnej jest opisanie różnic między grupami obiektów
- Celem predykcyjnej analizy dyskryminacyjnej jest klasyfikowanie obserwacji, o których nie wiadomo, do której grupy należą
- Może być traktowana jako rozszerzenie analizy wariancji
- Jeśli występują istotne statystycznie różnice w średnich wartościach cech obiektów pomiędzy grupami, to mogą być one wykorzystane do przewidywania przynależności do grupy

Liniowa Analiza Dyskryminacyjna (LDA)

- Zmienna objaśniana, wskazująca na przynależność do klas (grup), przyjmuje dwie albo większą (skończoną) liczbę wartości
- Poszukiwana jest taka kombinacja liniowa cech obiektów, która w optymalny sposób przypisze je do klas
- Procedura minimalizuje wariancję cech obiektu wewnątrz klasy i maksymalizuje wariancję pomiędzy klasami
- W tym celu poszukiwany jest kierunek wektora a

Czy można zidentyfikować drużynę piłkarską po jej wynikach? Analiza dyskryminacyjna

Błażej Popławski, Michał Sękowski

Plan raportu

- Cel badania
- Analiza dyskryminacyjna w piłce nożnej
- Opis zbioru danych
- Wykorzystane techniki
- Wyniki
- Podsumowanie

Cel badania

- Celem badania była identyfikacja drużyn piłkarskich na podstawie zebranych statystyk rozgrywanych meczy
- Wyniki analizy mogą być wykorzystane przez bukmacherów podczas ustalania kursów na wynik wydarzenia sportowego
- Klienci bukmacherów mogą wykorzystać wyniki budując swoje strategie
- Sztaby trenerskie zyskują narzędzie identyfikujące mocne i słabe strony przeciwnika

Analiza dyskryminacyjna w piłce nożnej

- Autorzy wskazują 4 podobne badania z ostatnich 15 lat
- W literaturze analiza dyskryminacyjna służyła do podziału na dwie lub trzy grupy
- Autorzy wskazują najczęściej wykorzystywanymi czynniki dyskryminacyjne
- oraz najczęściej wykorzystywaną miarę jakości modelu: procent prawidłowych klasyfikacji w tzw. tablicy klasyfikacyjnej

Opis zbioru danych

- Autorzy wskazali źródło danych oraz krótko opisali stworzone przez siebie narzędzie do pobrania danych (ang. webscrapping)
- Autorzy opisali procedurę wyboru obserwacji uzasadniając swoje postępowanie (7 sezonów, 11 drużyn, 266 meczy)
- Zaprezentowali i opisali podstawowe statystyki opisowe skonstruowanego zbioru

Opis zbioru danych

- Zmienna grupująca: nazwa drużyny
- Czynniki potencjalnie dyskryminujące:
  - Liczba bramek w pierwszej połowie
  - Liczba bramek na koniec meczu
  - % posiadania piłki
  - Liczba strzałów niecelnych
  - Liczba strzałów celnych
  - Liczba zablokowanych strzałów
  - Liczba rzutów różnych
  - % celnych podań
  - Liczba wygranych pojedynków w powietrzu
  - Liczba fauli
  - Liczba żółtych kartek
  - Liczba czerwonych kartek

Opis zbioru danych

- Angielska Premier League
  - 20 drużyn = 20 grup
  - 1 sezon = 380 meczów
  - Sezony 2014/15-2020/21
  - 7 sezonów = 2660 meczów
- Ograniczenie liczby drużyn do występujących w Premier League w każdym sezonie
  - 11 drużyn
  - 266 zdarzeń dla każdej drużyny

Opis zbioru danych. Drużyny

- Arsenal
- Chelsea
- Crystal Palace
- Everton
- Leicester City
- Liverpool
- Manchester City
- Manchester United
- Southampton
- Tottenham
- West Ham

Opis zbioru danych. Statystyki

- Statystyki opisowe wykorzystanych czynników dyskryminacyjnych
- Średnie i odchylenia standardowe zmiennych dyskryminujących z podziałem na drużyny
- Korelacje pomiędzy czynnikami dyskryminującymi

Wykorzystane techniki

W badaniu wykorzystano trzy techniki analizy dyskryminacyjnej:

- *liniową,*
- *kwadratową,*
- *k najbliższych sąsiadów „kNN”.*

- Są to techniki omawiane podczas zajęć więc ich szerzej nie opisywali

Wyniki

- Dla każdej metody analizy zestaw rezultatów:
- Funkcje dyskryminujące
- Korelacje między wartościami czynników a wartościami funkcji
- Zestawienie średnich wartości funkcji dyskryminacyjnych
- Ładunki standaryzowanych funkcji dyskryminacyjnych
- Tablica klasyfikacyjna

Podsumowanie

- Identyfikacja drużyn piłkarskich na podstawie zebranych statystyk rozgrywanych meczy jest możliwa
- Spośród rozpatrywanych metod najlepszą trafnością charakteryzowała się metoda kNN - ponad 75% obserwacji klasyfikowanych poprawnie

Ocena

- Praca ma prawidłowo przedstawiony problem i pytania badawcze
- Dane opisane bardzo dobrze
- Analiza przeprowadzona prawidłowo
- Zabrakło odniesienia wyników do literatury oraz wniosków odnoszących się do potencjalnych czynników dyskryminujących
- Zawartość wykresów i tabel mogłaby być bardziej czytelna

Analiza skupień

- Analiza skupień jest dziedziną eksploracji danych
- Jest to statystyczna metoda pozwalająca na znajdowanie grup podobnych obiektów w zbiorze danych
- Polega na dzieleniu wielowymiarowego zbioru danych na grupy (podzbiory) w taki sposób, by elementy w tej samej grupie były do siebie podobne, a jednocześnie jak najbardziej odmienne od elementów z pozostałych grup (podzbiorów)
- Analiza skupień znalazła wiele zastosowań w różnych dziedzinach, jak np. klasyfikacja dokumentów (analiza internetu), odkrywanie grup klientów o podobnych zachowaniach (marketing), czy wykrywanie oszustw kredytowych (banki)

Cele analizy

- Uzyskanie grup (skupień) jednorodnych obiektów, które ułatwiają wyodrębnienie ich cech
- Redukowanie dużej liczby danych pierwotnych do kilku podstawowych kategorii, które mogą być traktowane jako przedmioty dalszej analizy
- Ograniczenie czasu analizy, których przedmiotem będzie uzyskanie cech obiektów typowych
- Poznanie struktury analizowanych danych wielowymiarowych

Metody

- Analiza skupień jest metodą stworzoną raczej do formułowania hipotez na podstawie danych niż ich statystycznej weryfikacji
- Służy również opisowi wyodrębnionych podzbiorów, tzw "naturalnych skupień"
- Analiza niehierarchiczna traktująca grupy danych w sposób równoważny
- Analiza hierarchiczna zakładająca, że grupy danych są zagnieżdżone
- Współczesne metody takie jak np. DBSCAN (ang. Density-based spatial clustering of applications with noise)

Czy akcje charakteryzujące się podobnymi wynikami należą do jednego sektora - zastosowanie metod analizy skupień

Ewa Bogdanowicz, Dawid Lubiński, Dominika Miętek

Plan raportu

- Wprowadzenie
- Przegląd literatury
- Dane
- Metodologia
- Wyniki
- Podsumowanie i wnioski

Cel badania

- Celem badania była odpowiedź na pytanie: Czy ceny akcji spółek zachowują się w podobny sposób, jeżeli tak to czy grupy tych spółek należą do jednego sektora branżowego?
- Cena akcji zależy od czynników wewnętrznych i zewnętrznych
- Wewnętrzne to te, na które spółka ma wpływ, np. sytuacja finansowa, sposób zarządzania
- Na zewnętrzne spółka nie ma wpływu, np. sytuacja gospodarcza sektora, tempo rozwoju gospodarki, sytuacja makroekonomiczna

Literatura

- Autorzy omówili 3 podobne badania
- Zwracają uwagę na wykorzystywane metody i liczbę grup wybieraną w badaniach
- Zgodnie z literaturą tematu najlepsze wyniki zapewnia wykorzystanie metody k-średnich dla analizy niehierarchiznej oraz metody Warda dla analizy hierarchicznej

Dane

- Dane pobrano z bazy www.stooq.com, za okres 8.01.2021-18.01.2022.
- Uwzględniono spółki o kapitalizacji rynkowej powyżej 500 mln złotych
- Korekty danych
  - usunięcie spółek o więcej niż 10 brakujących obserwacjach
  - uzupełnienie braków z wykorzystaniem interpolacji liniowej
  - obliczenie logarytmicznych miesięcznych stóp zwrotu
- Baza finalna 101 spółek po 258 stóp zwrotu

Dane

- Baza finalna dla każdej spółki zawiera następujące czynniki (zmienne):
  - miesięczne logarytmiczne stopy zwrotu
  - miesięczne wariancje cen zamknięcia
  - wskaźnik cena/zysk
  - wskaźnik cena/wartośc księgowa

Dane. Analizowane sektory

- Autorzy wyróżnili sektory gospodarki (grupy) zgodnie z klasyfikacją wykorzystywaną przez GPW
  - chemia i surowce
  - dobra konsumpcyjne
  - finanse
  - handel i usługi
  - ochrona zdrowia
  - paliwa i energia
  - produkcja przemysłowa i budowlano-montażowa
  - technologie

Metodologia

- Analiza niehierarchiczna
  - Początkowa liczba skupień na podstawie kryteriów Calińskiego-Harabasza, metody Silhouette oraz wybrana subiektywnie
  - grupowanie metodą k-średnich z różnymi miarami odległości
- Analiza hierarchiczna
  - rozważane metody: najbliższego wiązania, średniego wiązania, pełnego wiązania i Warda dla różnych miar odległości
  - prezentacja dendrogramów
- Porównanie wyników uzyskanych metodami niehierarchicznymi i hierarchicznymi

Wyniki

- Dla obu metod autorzy zaprezentowali trzy warianty wyników dla odległości euklidesowej, miejskiej oraz maksymalnej
- W przypadku analiz niehierarchicznych autorzy wykorzystali metodę k-średnich
- W przypadku analiz hierarchicznych autorzy wykorzystali metodę Warda
- Dla każdego podziału autorzy scharakteryzowali cechy uzyskanych skupień

Wnioski

- Autorzy stwierdzili, iż na podstawie uzyskanych wyników nie można potwierdzić weryfikowanej hipotezy. Obie grupy metod prowadziły do niejednoznacznych wyników.
- Wyjątkiem były spółki z sektora bankowego oraz sektora gier
- Uzyskane prze autorów wyniki odbiegają od tych w literaturze

Ocena

- Praca ma prawidłowo sformułowaną hipotezę badawczą
- Dane opisane bardzo dobrze
- Analiza przeprowadzona prawidłowo, mogłaby być opisana w sposób bardziej szczegółowy
- Zabrakło odniesienia poszczególnych wyników do literatury. Ogólny komentarz to za mało. Autorzy nie wykorzystali prawidłowo zgromadzonej literatury (8 pozycji)
- Praca czytelna, brak uchybień formalnych i edytorskich