Poznań, 14 grudnia 2002

Case Study 2 – Analiza skupień

Celem ćwiczenia jest przeprowadzenie procesu grupowania / analizy skupień dla jednego z wybranych zbiorów danych (tj. dostarczonych przez prowadzącego). W trakcie realizacji tego case study należy także zapoznać się z różnymi metodami grupowania i różnym doborem parametrów dla każdej ze stosowanych metod.

W przypadku stosowania pakietu *Statistica* – proszę stosować zaimplementowane tam algorytmy aglomeracyjne i *k* – średnich.

Uwaga: dostarczane dane mają dość surowy format. Należy zawsze **zbadać poprawność** dostarczonych danych. Jeśli poszczególne atrybuty/cech zdefiniowane są na różnorodnych skalach pomiarowych lub skalach liczbowych o różnym zakresie dziedzin, zaleca się wykonanie **przetwarzania wstępnego** – w szczególności konieczne jest dokonanie **normalizacji** lub **standaryzacji**. Ponadto definicje zbiorów atrybutów/cech zawierają nadmiarowe atrybuty – zwłaszcza dla danych charakteryzujących zróżnicowanie poziomu życia w poszczególnych województwach Polski.

Zalecane jest najpierw wykonanie grupowania **metodą aglomeracyjną**, następnie po analizie różnych dendrogramów można wykonać **grupowanie metodą *k*-średnich** (dla *k* wynikającego z poprzedniej analizy i własnej interpretacji charakterystyki problemy).

W ramach każdego z algorytmów grupowania konieczne jest zbadanie różnych parametrów i wpływu ich na wynik końcowy (hierarchię i budowę skupień oraz konkretny opis utworzonych skupień). Przy czym nie jest konieczne sprawdzanie wszystkich kombinacji różnych parametrów, ale sprawdzenie tylko kilku najbardziej dogodnych dla danego problemu.

Należy w sprawozdaniu prowadzić dyskusję **dobieranych parametrów** (zwłaszcza dla metody aglomeracyjnej) i uzasadnić wybór końcowy.

Przykładowo dla grupowania aglomeracyjnego wykonaj hierarchizacje przy rożnych definicjach odległości, i sposobach aglomeracji. Zastanów się:

- Czy miara odległości wpływa znacząco na wyniki?
- Zilustrować obliczenia wykresami drzewkowymi.
- Zidentyfikuj grupy przykładów obiektow podobnych.
- Zidentyfikuj ewnatulane obserwacje odstające i znajdź przyczyny ich „nietypowosci”.
- Przeprowadź podobna analize dla wybranych podzbiorow cech (atrybutów, zmiennych) – zalecane, gdyż niektóre dane mają nadmiarowe zbiory cech.

(Uwaga: Program STATISTICA nie potrafi przeprowadzac analizy dla pojedynczych zmiennych!).

Podobne pytania możesz sobie postawić dla grupowania *k*-średnich.

Każdorazowo należy również zwracać uwagę na interpretację wyniku końcowego, np.:

- Dla metody aglomeracyjnej dokładnie analizować strukturę drzewa dendrogramu (oceniając czy drzewa ma właściwe „zrównoważenie” liczebności skupień, lub czy pojedyncze skupienia są pożądane),
- Warto analizować także wykres odległości wiązanie względem kolejnych iteracji/etapów algorytmu – na jego podstawie i obserwacji drzewa hierarchii należy podjąć decyzje, co do możliwej liczby skupień.

- Starać się analizować opis skupień za pomocą wartości średnich, odchyleń i innych parametrów,
- Spróbować podać własną interpretację przynależności podobnych obiektów do skupień

Pamiętaj, że należy ocenić czy wszystkie z rozpatrywanych zmiennych są konieczne i pożyteczne w procesie analizy – zbędne wyeliminować.
Całość przebiegu analizy i interpretacji doboru parametrów oraz wyników końcowych powinno być zawarte w sprawozdaniu końcowym.

Analizie podlegają następujące zbiory danych:

1. jedjewro.sta

2. wojew1.xls

3. wojew2.xls

Załączniki:

1. „Wyciąg” z helpu do modułu analizy skupień
2. Krótka charakterystyka zbiorów danych.

Załącznik nr 1 do case study nr 2

Krótka charakterystyka zbiorów danych przeznaczonych dla przećwiczenia analizy skupień w case study nr 2

Dane „jedjewro"

Dane zawierają informacje o spożyciu różnych produktów żywnościowych w poszczególnych krajach europejskich. Dokładnej dla każdego z wybranych produktów podaje się wartość oszacowanej konsumpcji protein obliczanej na osobę i dzień – może mieć to znaczenie przy analizie sposobu odżywiania się mieszkańców różnych krajów europejskich..
Rozważa się następujące zmienne/produkty:
mięso wołowe,
wieprzowina i drób,
jaja,
mleko,
ryby,
produkty mączne,
pokarmy o dużej zawartości skrobi,
rośliny strączkowe, orzechy, nasiona roślin oleistych,
owoce i warzywa

Celem analizy jest zidentyfikowanie regionów Europy, które charakteryzują się wewnętrznym podobieństwem ze względu na spożycie powyższych produktów i udział w nich protein. Ponadto dokonaj opisu charakterystyki tych wyróżnionych regionów, aby zidentyfikować, jakie są charakterystyczne wartości różnicujące regiony pomiędzy sobą.

Źródło danych to praca Greenacre (1984); z Weber, (1973). Agrarpolitik im Spannungsfeld der Internationlen Ernährungspolitik. Kiel, Institut für Agrarpolitik und Marktlehre.
Zwróć uwagę, że struktura spożycia jest charakterystyczna dla krajów europejskich przed kilkudziesięciu laty – obecnie mogła ulec zmianie.

Dane „województwa"

Dane zawierają informacje o **przestrzennym zróżnicowaniu poziomu życia** w Polsce na początku lat **90**tych. Zgodnie z badaniami statystycznymi poziom i warunki życia ludności w Polsce są silnie zróżnicowane przestrzennie, co jest powiązane z ogólnym rozwojem społeczno-gospodarczym poszczególnych **województw**, poziomem wytwarzanego produktu krajowego brutto na 1 mieszkańca, poziomem uprzemysłowienia, poziomem wykształcenia, zamożnością przeciętnego mieszkańca, standardem jego życia, stopniem urbanizacji, a także funkcjonowaniem wielkich aglomeracji miejskich na terenia województw. Inne czynniki mogą obejmować dostępność określonych usług, szkolnictwa na odpowiednim poziomie, dóbr kultury, różnych sposobów wykorzystania czasu wolnego, a także aspekty ekologiczne i ochrony środowiska. Podstawową jednostką, którą charakteryzuje się różnymi atrybutami jest województwo – w podziale na 49 województw przed reformą administracyjną.

Celem analizy jest zidentyfikowanie regionów Polski (grup pewnych województw), które charakteryzują się wewnętrznym podobieństwem ze względu na poziom życia w nich. Ponadto dokonaj opisu charakterystyki tych wyróżnionych regionów, aby zidentyfikować, jakie są charakterystyczne wartości różnicujące regiony pomiędzy sobą.

Dostępne są dwa rodzaje zbiorów danych:

1. Charakteryzuje województwa bardziej z punktu widzenia parametrów ekonomicznych, wskaźników finansowych. jak i poziomu rozwoju określonych aspektów przemysłowo-usługowych i poziomu szkolnictwa?
2. W mniejszym stopniu charakteryzuje województwa współczynnikami finansowymi; dodatkowo zawiera informacje na temat stopy życiowej, zagrożeń cywilizacyjnych, stylu i standardu życia.

W trakcie analizy pamiętaj, że dane dotyczą okresu początku lat 90-tych, co związane jest z użyciem być może innych jednostek i zakresu pomiarów niż dostępne obecnie – np. w przypadku jednostek pieniężnych, poziomu bezrobocia,...
Zwróć uwagę, że podane definicje zbiorów atrybutów/cech mogą zawierać nadmiarowe informacje (liczba wskaźników jest potencjalnie zbyt liczna), oraz niektóre atrybuty mogą być nie w pełni zdefiniowane.

Uwagi do przebiegu analizy skupień – wybrane fragmenty z „Help-u" programu Statistica

Wprowadzenie do metod aglomeracyjnych - Hierarchiczne drzewo

Rozważmy poziomy hierarchiczny wykres drzewkowy rozpoczynając od lewej strony wykresu, gdzie każdy obiekt stanowi swoją własną klasę. Wyobraźmy sobie teraz, że bardzo małymi krokami „osłabiamy" nasze kryterium tego, na ile jest on lub nie jest wyjątkowy. Innymi słowy, obniżamy próg stanowiący o decyzji przypisania dwóch lub więcej obiektów do tego samego skupienia. Tym sposobem wiążemy ze sobą coraz to więcej obiektów i agregujemy je w coraz to większe skupienia elementów coraz bardziej różniących się od siebie. W końcu, na ostatnim etapie, wszystkie obiekty zostają ze sobą połączone.
Na wykresach tych na osi poziomej odłożone są odległości aglomeracyjne (w pionowych wykresach sopelkowych odległość aglomeracyjna odkładana jest na osi pionowej). Zatem przy każdym węźle na wykresie (gdzie uformowało się nowe skupienie) **możemy odczytać odległość, przy której odpowiednie elementy** zostały powiązane ze sobą tworząc nowe pojedyncze skupienie. Jeśli dane mają wyrazistą „strukturę" w tym sensie, że istnieją skupienia podobnych do siebie obiektów, to często struktura ta znajdzie odbicie na hierarchicznym drzewie w postaci oddzielnych gałęzi. Pomyślna analiza przy pomocy metody łączenia daje możliwość wykrywania skupień (gałęzi) i ich interpretacji.

Wyniki aglomeracji - Hierarchiczny wykres drzewkowy

Przycisk **Poziomy hierarchiczny wykres drzewkowy:** Naciśnięcie tego przycisku spowoduje utworzenie poziomego diagramu drzewkowego, który przedstawia następstwo grupowania obiektów. Informacje na temat interpretacji diagramu drzewkowego znajdują się w części Wprowadzenie do analizy skupień.

Przycisk **Pionowy wykres sopelkowy**: Naciskamy ten przycisk, aby utworzyć pionowy diagram drzewkowy (odległości wiązania odłożone są na osi pionowej). I znów polecamy odwołanie się do części Wprowadzenie do analizy skupień, gdzie znajdują się informacje na temat interpretacji diagramu drzewkowego.

Prostokątne gałęzie: Na obu typach wykresów drzewkowych (patrz powyżej) mamy możliwość wyboru wyświetlania albo prostokątnych gałęzi (zaznaczamy opcję), albo ukośnych gałęzi (anulujemy zaznaczenie opcji). Drugi format może podnieść czytelność diagramu w przypadku rozwiązań ze „zrównoważonymi" strukturami łączenia.

Skaluj drzewo do odl_wiązania/odl_maks*100: Po wybraniu tej opcji, drzewo zostanie przeskalowane do skali standaryzowanej (tzn. odległość wiązania/odległość maksymalna*100). W przeciwnym wypadku, gdy nie wybierzemy tej opcji, skala będzie oparta na odległości wiązania wybranej w Panelu początkowym.

UWAGA: W zależności od bieżącego ustawienia domyślnego rozmiaru czcionki dla wartości skali (patrz Domyślne opcje skal), etykiety pozycji (przypadków lub zmiennych) na wykresie drzewkowym mogą się nakładać lub część (np. co druga lub co trzecia) z nakładających się etykiet może zostać pominięta (patrz Filtry); w takich wypadkach klikamy etykiety na wykresie i zmniejszamy rozmiar czcionki.

Wyniki aglomeracji - Przebieg aglomeracji

Kliknięcie tego przycisku przywoła arkusz wyników z opisem przebiegu procesu aglomeracji. Pierwsza kolumna arkusza wyników będzie zawierać odległości wiązań, na których zostały uformowane odpowiednie skupienia (wskazane w odpowiednich wierszach), a każdy wiersz zawiera nazwy obiektów (przypadków lub zmiennych), które formują dane skupienie.

Wprowadzenie do metod aglomeracyjnych - Metody amalgamacji lub wiązania

Na pierwszym etapie, gdy każdy obiekt reprezentuje swoje własne skupienie, odległości między tymi obiektami definiuje się przy pomocy wybranej miary odległości. Jak jednak określić odległości między nowymi skupieniami, które powstaną z powiązanych obiektów? Innymi słowy, potrzebujemy zasady wiązania lub aglomeracji, która określi, kiedy dwa skupienia są dostatecznie podobne, aby można je było połączyć. Istnieje kilka możliwości: na przykład, moglibyśmy powiązać ze sobą dwa skupienia, gdy dowolne dwa obiekty z tych dwóch skupień znajdują się w mniejszej odległości niż odpowiednia odległość wiązania. Innymi słowy, aby określić odległości między skupieniami, wykorzystamy „najbliższych sąsiadów" między skupieniami; metoda ta nosi nazwę pojedynczego wiązania [single linkage]. W wyniku zastosowania tej metody powstają skupienia typu „włóknistego", co oznacza, że są one „przymocowane do siebie" tylko przez pojedyncze obiekty, które leżą najbliżej siebie. Alternatywnie, możemy wykorzystać sąsiadów, którzy są najbardziej od siebie oddaleni; ta metoda nosi nazwę pełnego wiązania [complete linkage]. Istnieje wiele innych zasad wiązania podobnych do zaproponowanych tutaj, a moduł analizy skupień oferuje duży ich wybór.

Metoda pojedynczego wiązania (najbliższego sąsiedztwa) [Single linkage (nearest neighbor)**]**. Jak to zostało opisane powyżej, w metodzie tej odległość między dwoma skupieniami jest określona przez odległość między dwoma najbliższymi obiektami (najbliższymi sąsiadami) należącymi do różnych skupień. Zgodnie z tą zasadą obiekty formują skupienia łącząc się w sznur, a wynikowe skupienia tworzą długie „łańcuchy".

Metoda pełnego wiązania (najdalszego sąsiedztwa) [Complete linkage (furthest neighbor)**]**. W tej metodzie odległość między skupieniami jest zdeterminowana przez największą z odległości między dwoma dowolnymi obiektami należącymi do różnych skupień (tzn. „najdalszymi sąsiadami"). Metoda ta zwykle zdaje egzamin w tych przypadkach, kiedy obiekty faktycznie formują naturalnie oddzielone „kępki". Metoda ta nie jest odpowiednia, jeśli skupienia są w jakiś sposób wydłużone lub mają naturę „łańcucha".

Metoda średnich połączeń [Unweighted pair-group average]. W metodzie tej odległość między dwoma skupieniami oblicza się jako średnią odległość między wszystkimi parami obiektów należących do dwóch różnych skupień. Metoda ta jest efektywna, gdy obiekty formują naturalnie oddzielone „kępki", ale zdaje także egzamin w przypadku skupień wydłużonych, mających charakter „łańcucha". Zwróćmy uwagę, że Sneath i Sokal (1973) na określenie tej metody wprowadzili skrót UPGMA (unweighted pair-group method using arithmetic averages).

Metoda średnich połączeń ważonych [Wieghted pair-group average]. Jest to metoda identyczna jak metoda średnich połączeń, z tym wyjątkiem, że w obliczeniach uwzględnia się wielkość odpowiednich skupień (tzn. liczbę zawartych w nich obiektów) jako wagę. Zatem metoda ta (raczej niż poprzednia) powinna być stosowana wtedy, gdy podejrzewamy, że liczebności skupień są wyraźnie nierówne. Sneath i Sokal (1973) na określenie tej metody wprowadzili skrót WPGMA (weighted pair-group method using arithmetic averages).

Metoda środków ciężkości [Unweighted pair-group centroid]. Środek ciężkości skupienia jest średnim punktem w przestrzeni wielowymiarowej zdefiniowanej przez te wymiary. W metodzie tej, odległość między dwoma skupieniami jest określona jako różnica między środkami

ciężkości. Sneath i Sokal (1973) na oznaczenie tej metody stosują skrót UPGMC (unweighted pair-group method using the centroid average).

Metoda ważonych środków ciężkości (mediany) [Weighted pair-group centroid]. Jest to metoda identyczna jak poprzednia, z tym wyjątkiem, że w obliczeniach wprowadza się ważenie, aby uwzględnić różnice między wielkościami skupień (tzn. liczbą zawartych w nich obiektów). Zatem, metoda ta jest lepsza od poprzedniej w sytuacji, gdy istnieją (lub podejrzewamy, że istnieją) znaczne różnice w rozmiarach skupień. Sneath i Sokal (1973) na oznaczenie tej metody stosują skrót WPGMC (weighted pair-group method using the centroid average).

Metoda Warda. Ta metoda różni się od wszystkich pozostałych ponieważ do oszacowania odległości między skupieniami wykorzystuje podejście analizy wariancji. Krótko mówiąc metoda ta zmierza do minimalizacji sumy kwadratów dowolnych dwóch skupień, które mogą zostać uformowane na każdym etapie. Szczegóły na temat tej metody znajdują się w: Ward (1963). Ogólnie, metoda ta jest traktowana jako bardzo efektywna, chociaż zmierza do tworzenia skupień o małej wielkości.

Przegląd dwóch innych metod grupowania znajduje się w częściach: Grupowanie obiektów i cech oraz Grupowanie metodą k-średnich.

Grupowanie metodą k-średnich.

Ogólna zasada
Ta metoda grupowania różni się znacznie od metod Aglomeracji i Grupowania obiektów i cech. Załóżmy, że sformułowaliśmy już hipotezę na temat liczby skupień naszych przypadków lub zmiennych. Możemy „powiedzieć" komputerowi, aby uformował dokładnie 3 skupienia, które będą tak różne, jak to tylko możliwe. Temu typowi problemu badawczego odpowiada algorytm grupowania metodą k-średnich. Ogólnie, przy pomocy metody k-średnich zostanie utworzonych k różnych możliwie odmiennych skupień.

Przykład. W przykładzie dotyczącym sprawności fizycznej (patrz Grupowanie obiektów i cech), badacz danych medycznych może się na podstawie doświadczenia klinicznego „domyślać", że pacjenci z chorobami serca wpadną do trzech różnych kategorii ze względu na sprawność fizyczną. Może być ciekawy, czy ta intuicja może zostać określona ilościowo, to znaczy, czy zgodnie z oczekiwaniem, analiza skupień metodą k-średnich miar sprawności fizycznej faktycznie utworzyłaby trzy skupienia pacjentów. Jeśli tak, średnie różnych miar sprawności fizycznej dla każdego skupienia reprezentowałyby ilościowy sposób wyrażenia hipotez lub intuicji badacza (tzn. pacjenci w skupieniu 1 są wysoko według miary 1, nisko według miary 2 itd.).

Obliczenia. Z punktu widzenia obliczeń, można tę metodę traktować jako „odwrotność" analizy wariancji (ANOVA). Program rozpocznie od k losowych skupień, a następnie będzie przenosić obiekty między tymi skupieniami mając na celu (1) minimalizację zmienności wewnątrz skupień i (2) maksymalizację zmienności między skupieniami. Jest to analogiczne do „odwrotności" analizy wariancji w tym sensie, że test istotności w analizie wariancji szacuje zmienność międzygrupową w stosunku do zmienności wewnątrzgrupowej, jeśli liczymy test istotności dla hipotezy, że średnie w grupach różnią się między sobą. W grupowaniu metodą k-średnich program stara się przenosić obiekty (np. przypadki) do i z grup (skupień), aby otrzymać najbardziej istotne wyniki analizy wariancji. (Ponieważ, obok innych rezultatów, wyniki analizy wariancji są częścią standardowego zestawienia wyników analizy opartej na grupowaniu metodą k-średnich, użytkownik może odwołać się do części ANOVA/MANOVA, aby dowiedzieć się czegoś więcej na temat tej metody.)

Interpretacja wyników. Zazwyczaj w wyniku analizy grupowania metodą k-średnich badamy średnie dla każdego skupienia w każdym wymiarze, aby oszacować, na ile nasze k skupienia są

od siebie różne. W sytuacji idealnej otrzymalibyśmy bardzo różne średnie dla większości, jeśli nie wszystkich wymiarów wprowadzonych do analizy. Wielkość statystyki F pochodzącej z analizy wariancji wykonanej w każdym wymiarze jest wskaŹnikiem tego, na ile dobrze dany wymiar dyskryminu

Grupowanie obiektów i cech - Wprowadzenie

Poprzednio omawialiśmy tę metodę w kategoriach „obiektów", które mają zostać pogrupowane (patrz Aglomeracja). We wszystkich innych typach analiz w programie STATISTICA pytanie badawcze jest zwykle wyrażone w kategoriach przypadków (obserwacji) lub zmiennych. Okazuje się, że grupowanie ich może doprowadzić do ciekawych wyników. Na przykład, wyobraŹmy sobie studium, w którym badacz w naukach medycznych zgromadził dane na temat różnych miar sprawności fizycznej (zmienne) dla próby pacjentów z chorobami serca (przypadki). Badacz może chcieć poklasyfikować przypadki (pacjentów), aby wykryć skupienia pacjentów o podobnych syndromach. Jednocześnie badacz może chcieć poklasyfikować zmienne (miary sprawności), aby wykryć skupienia miar, które dotyczą podobnych zdolności fizycznych. W module analizy skupień możemy wybrać zarówno grupowanie przypadków jak i zmiennych.

Grupowanie obiektów i cech

Po dyskusji w powyższym akapicie dotyczącej tego, czy grupować przypadki, czy zmienne, można spytać dlaczego by nie grupować jednych i drugich jednocześnie? Moduł analizy skupień zawiera procedurę grupowania obiektów i cech, która służy właśnie do tego. Grupowanie obiektów i cech przydaje się w (stosunkowo rzadkich) okolicznościach, gdy oczekujemy, że zarówno przypadki, jak i zmienne jednocześnie przyczyniają się do odkrywania sensownych układów skupień. Na przykład, wracając do tego samego przykładu, badacz w naukach medycznych może chcieć zidentyfikować skupienia pacjentów, którzy są podobni ze względu na poszczególne skupienia podobnych miar sprawności fizycznej. Trudność w interpretacji takich wyników może brać się stąd, że podobieństwa między różnymi skupieniami mogą odnosić się do (lub wynikać z) nieco innych podzbiorów zmiennych. Zatem wynikowa struktura (układ skupień) z natury nie jest homogeniczna. Z początku może się to wydawać dość niejasne i faktycznie, porównując z innymi opisanymi metodami grupowania (patrz Aglomeracja i Grupowanie metodą k-średnich), grupowanie obiektów i cech jest prawdopodobnie wykorzystywane najrzadziej. Niektórzy badacze jednak wierzą, że metoda ta stanowi mocne narzędzie eksploracyjnej analizy danych (szczegółowy opis tej metody znajduje się w: Hartigan, 1975).