SI - SPRAWOZDANIE LAB NR 9

Maciej Budzowski, dzienne, grupa L5 01/06/2020

Link do pliku .tex na platformie OverLeaf: OVERLEAF

Analiza danych przy użyciu sieci Kohonena

Użyty zbiór danych

Analiza została wykonana na zbiorze ***Wholesale customers***, linki do źródeł zamieszczam poniżej:
DANE (PLIK) (link prowadzi do oryginalnego pliku)
DANE (STRONA) (link prowadzi do strony datasetu)

Opis/temat zbioru: Wedle strony datasetu: “Zestaw danych odnosi się do klientów dystrybutora hurtowego. Obejmuje roczne wydatki w jednostkach monetarnych (m.u.) na różne kategorie produktów”

Autor:
Margarida G. M. S. Cardoso, margarida.cardoso ’@’ iscte.pt, ISCTE-IUL, Lisbon, Portugal

Parametry użytego zbioru:
- liczba obserwacji - 440
- liczba cech, atrybutów - 8

Lista cech w zbiorze (wedle strony):

- 1) FRESH: roczne wydatki (m.u.) na świeże produkty (Liczbowe);
- 2) MILK: roczne wydatki (m.u.) na mleczne produkty (Liczbowe);
- 3) GROCERY: roczne wydatki (m.u.) na spożywcze produkty (Liczbowe);
- 4) FROZEN: roczne wydatki (m.u.) na mrożone produkty (Liczbowe)

- 5) DETERGENTS-PAPER: roczne wydatki (m.u.) na detergenty i papierowe produkty (Liczbowe)
- 6) DELICATESSEN: roczne wydatki (m.u.) na delikatesy (Liczbowe);
- 7) CHANNEL: klienci (kanał sprzedaży) - Horeca (Hotel/Restaurant/Cafe) lub handel detaliczny (Nominalne)
- 8) REGION: Region - Lisbon, Oporto or Other (Nominalne)

Przygotowanie danych

W ramach przygotowania danych do pracy nie musiałem dokonywać zmian bezpośrednio w oryginalnym pliku. Natomiast zgodnie z metodą z odbytych ćwiczeń analiza została dokonana bez nazw klas.

Kod

Link do projektu ze skryptem: GITHUB

Sieci Kohonena tworzyłem przy użyciu skryptu języka R. Opiera się on na kodzie z odbytych ćwiczeń.

Proces analizy przebiega następująco (w kolejności):

- Wczytujemy plik z danymi.
- Pozbywamy się niepotrzebnych kolumn z nazwami klas (region i channel).
- Dane przekształcamy do macierzy.
- Definiujemy siatkę SOM.
- Uruchamiamy algorytm trenowania SOM
- Rysujemy wykresy i dokonujemy obserwacji
- Pozostawiamy komentarz

Obliczenia wykonywały się szybko, więc w tym przypadku nie było potrzeby automatyzacji procesu.

Proces wykonałem (tak jak na ćwiczeniach) w kilku próbach, przy użyciu różnych wielkości siatki SOM. Ostatecznie do próby wyciągnięcia informacji z danych zdecydowałem się (ze względu na ilość danych i otrzymywane wykresy) na siatkę 4x4.

Wyniki

Poniżej przedstawiam pojedyncze wykresy dla otrzymanych danych.

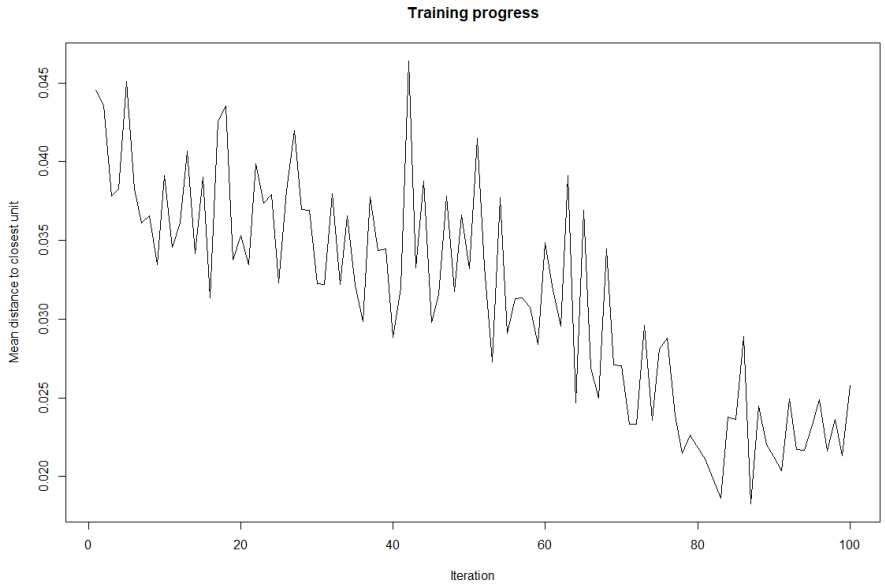

Figure 1: wykres typu: *changes*

Wykres typu *changes* przedstawia średnią odległość między neuronami.

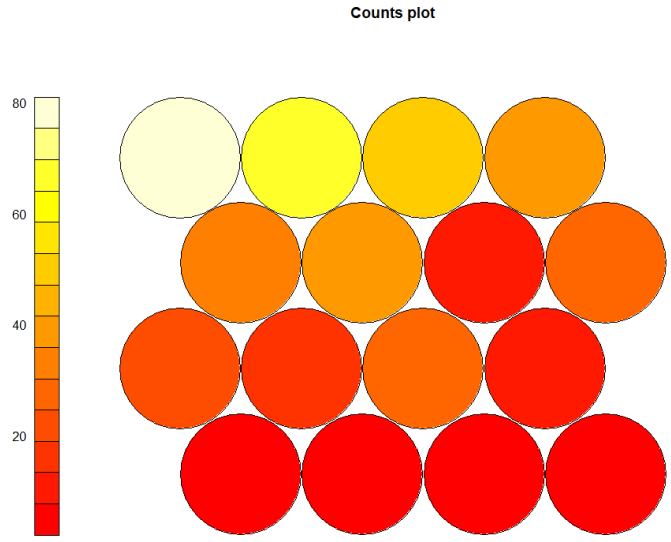

Figure 2: wykres typu: *count*

Wykres typu *count* - każde kółko to jeden neuron, rysunek interpretuje liczbę neuronów, które są blisko danego obiektu.

Figure 3: wykres typu: *dist.neighbours*

Wykres typu *dist.neighbours* - interpretuje odległość do najbliższych sąsiadów. (obiektów)

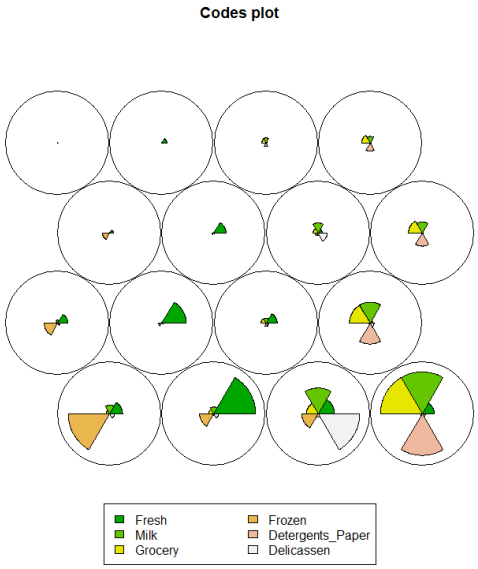

Figure 4: wykres typu: *codes*

Wykres typu *codes* - interpretuje udziały procentowe wartości parametrów.

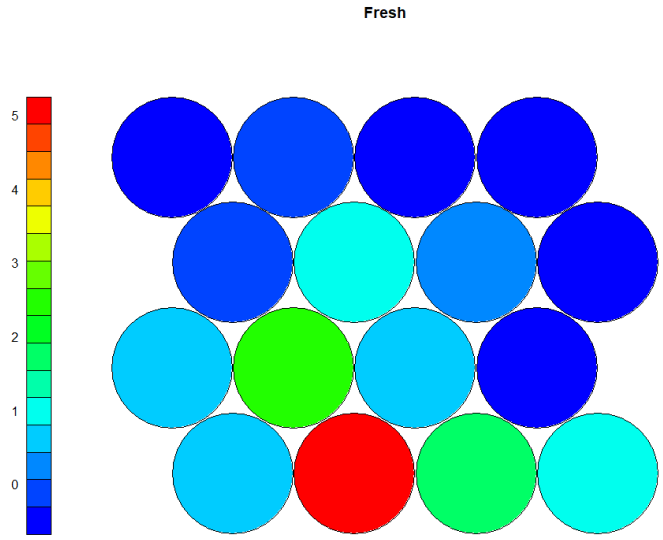

Figure 5: wykres dla parametru *Fresh*

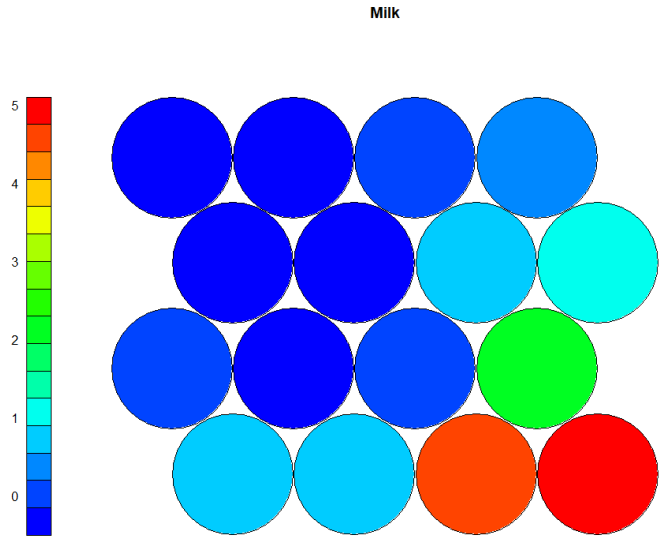

Figure 6: wykres dla parametru *Milk*

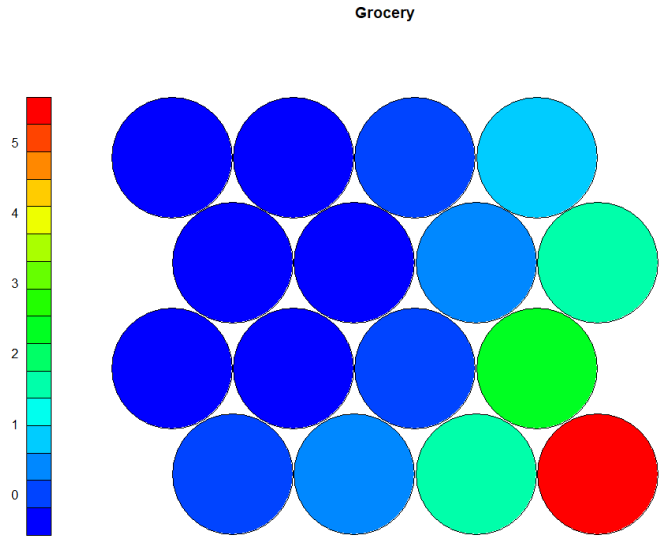

Figure 7: wykres dla parametru *Grocery*

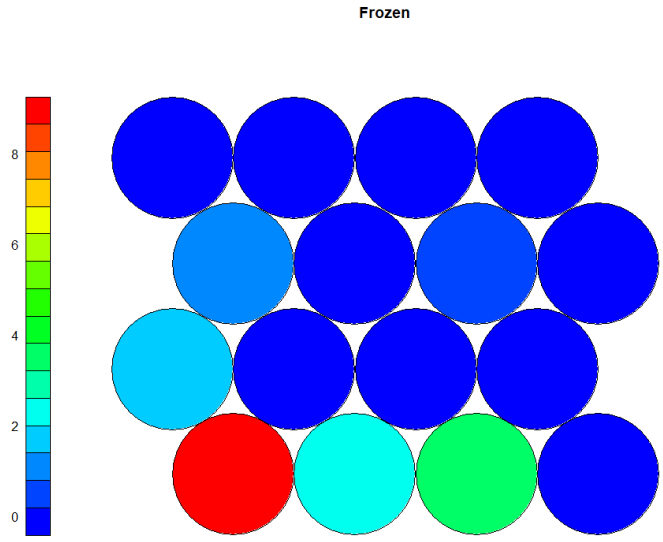

Figure 8: wykres dla parametru *Frozen*

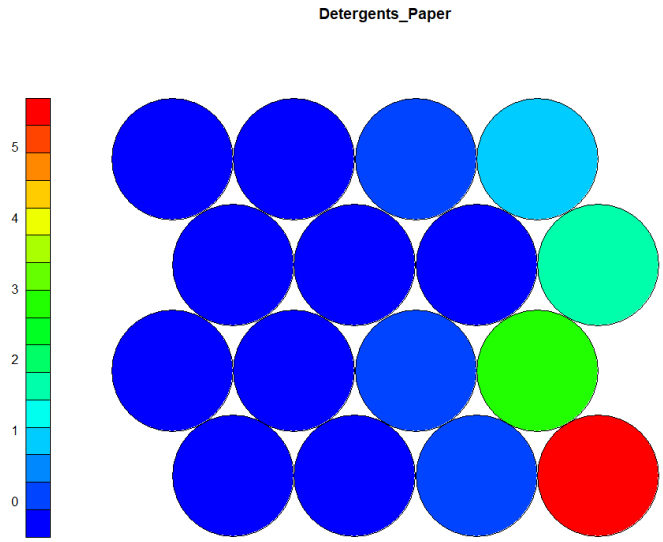

Figure 9: wykres dla parametru *Detergents and Paper*

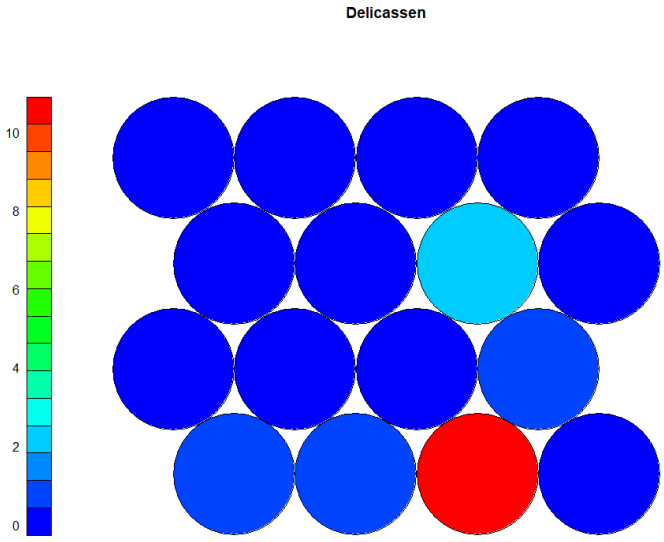

Figure 10: wykres dla parametru *Delicassen*

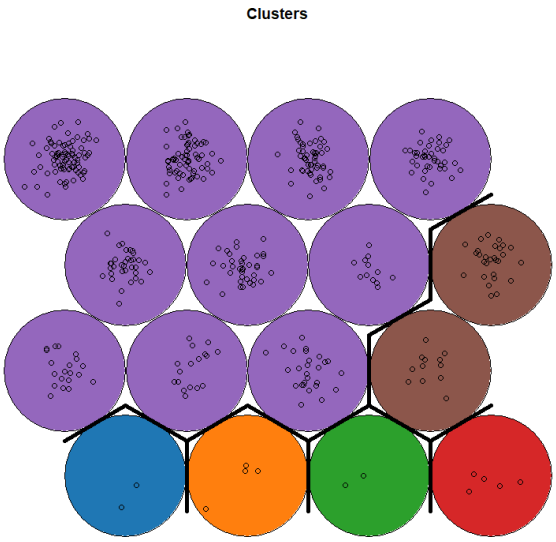

Figure 11: grupy, które zostały uzyskane metodą klasteryzacji hierarhicznej

Wnioski, obserwacje

Wykresy nałożyłem na siebie przy użyciu programu GIMP odpowiednio dopasowując wartość parametru "krycie".

Porównując ze sobą wykresy w poszukiwaniu korelacji znalazłem 2 możliwości, które są - moim zdaniem - warte zanotowania. Moim zdaniem, że najbardziej przydatnym wykresem, jest wykres typu *codes*.

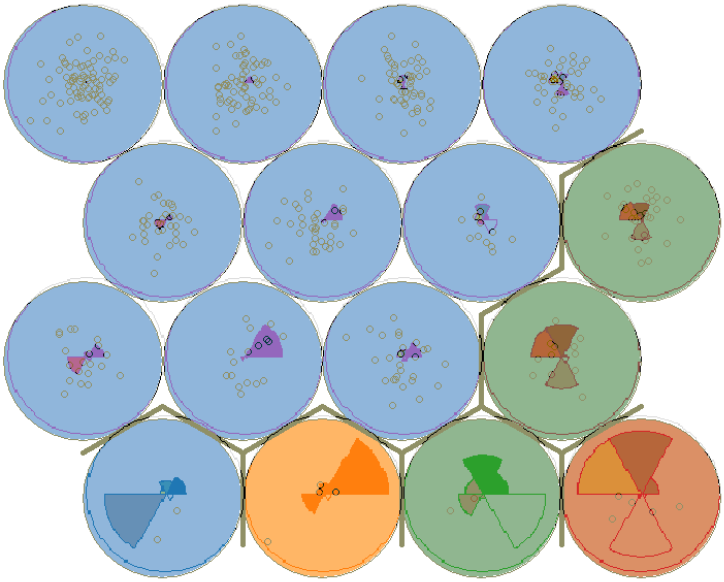

Figure 12: zmiksowane wykresy *cluster* i *codes*

W przypadku rysunku 12 mamy zmiksowane wykresy *cluster* i *codes*. Jako, że dane dotyczą wydatków na różne rodzaje produktów przez klientów na określonym obszarze, możemy dojść do następujących wniosków:

- Mały odsetek obiektów stanowi znaczną część wydatków.
- W grupach, które zostały utworzone metodą klasteryzacji hierarchicznej można zauważyć tendencję do kupowania określonego rodzaju produktu w dużej ilości (czerwona - grocery/detergents, zielona - milk/delicassen, pomarańczowa - fresh).
- Grupa niebieska stanowi większość obiektów, natomiast stanowi ona stosunkowo mały procent w całości wydatków. Można przewidywać, że mamy tutaj do czynienia z małymi biznesami.
- Grupa czerwona jest "elitą" - jest to kilka obiektów, które stanowią duży procent w całości wydatków. Można przewidywać, że są to duże przesiębiorstwa.

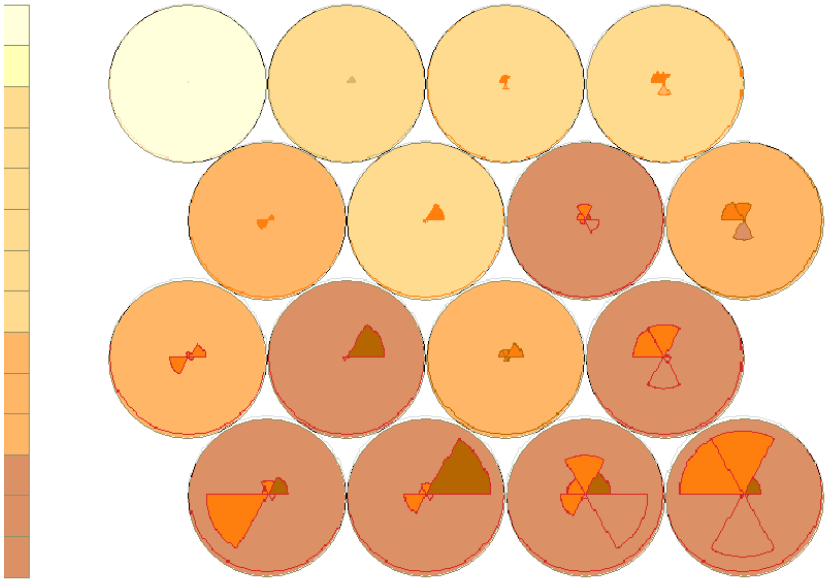

Figure 13: zmiksowane wykresy *codes* i *count*

W przypadku rysunku 13 mamy zmiksowane wykresy *codes* i *count*. Znajdujemy tutaj dodatkowe potwierdzenie poprzednich obserwacji. Stosunkowo mała ilość obiektów stanowi wysoki procent w całości wydatków. Natomiast obiekty stanowiące większość (od środka do lewego górnego rogu) stanowi mały procent w całości wydatków.