Projekt MOW - porównanie sieci SOM z klasycznymi metodami grupowania

Artur M. Brodzki, Adam Małkowski

15 listopada 2017

1 Wstęp - interpretacja tematu projektu

W ramach projektu dokonamy porównania jakości grupowania uzyskanego przy pomocy samo-organizujących się map (sieci SOM) oraz bardziej klasycznych metod grupowania: metody k-średnich, metody DBSCAN oraz metody najdalszego sąsiedztwa (ang. *comlete-linkage clustering*). Przy wykorzystaniu dostępnych powszechnie zbiorów danych zamierzamy wyznaczyć klastry metodami klasycznymi oraz za pomocą sieci SOM, a następnie porównać jakość tych klastrów z użyciem kilku powszechnie wykorzystywanych w tym celu metryk.

2 Wykorzystywane algorytmy

Samo-organizujące się mapy zrealizujemy w środowisku R za pomocą pakietów *som*, *kohonen*. W problemie grupowania ważną kwestią jest przyjęta definicja klastra. Nie ma jednej definicji klastra uznanej za standardową, a różne algorytmy przyjmują na swój użytek różne definicje:

1. Sieci SOM mogą grupować na dwa sposoby. Ponieważ każdy neuron sieci SOM odpowiada pewnemu zapamiętanemu wzorcowi danych, można założyć, że każdy taki neuron - wzorzec definiuje pewien wytworzony przez sieć SOM klaster danych. Ponieważ liczba neuronów - klastrów jest z góry określona, takie podejście charakteryzuje się znacznym podobieństwem do algorytmu k-średnich (w przypadku granicznym, gdy promień sąsiedztwa sieci SOM jest równy 0, obie metody są równoważne).

   Drugie podejście do grupowania na sieciach SOM wykorzystuje fakt, że neurony tej sieci są zanurzone w przestrzeni euklidesowej (zazwyczaj dwuwymiarowej). Dynamika sieci wymusza rozmieszczanie podobnych sobie wzorców blisko siebie. Dzięki temu, same neurony - wzorce mogą być obiektem grupowania. Na takich neuronach można uruchomić jeden z klasycznych algorytmów grupowania i porównać jakość uzyskanych klastrów. Taka procedura pozwala też zweryfikować skuteczność sieci SOM do odnajdowania powiązanych ze sobą wzorców w danych.

2. Podejście centroidowe - klaster jest geometrycznym środkiem zbioru bliskich sobie punktów. W celu analizy podejścia centroidowego wykorzystamy algorytm k-średnich.

3. Podejście hierarchiczne - klastry są wyznaczane na różnych poziomach jako złączone elementy z poziomu niższego. Do analizy podejścia hierarchicznego wykorzystamy algorytm najdalszego sąsiedztwa.

4. Podejście gęstościowe - klastry to zagęszczone skupiska położonych blisko siebie punktów. Granice między klastrami leżą w obszarach o mniejszej gęstości punktów. Analizę podejścia gęstościowego przeprowadzimy na algorytmie DBSCAN.

3 Przygotowanie danych

Do analizy metod grupowania wykorzystamy 3 spośród zbiorów danych dostępnych powszechnie zbiorów danych:

1. Car Evaluation Data Set - zbiór liczący 1728 elementów opisanych parametrami słownikowymi http://archive.ics.uci.edu/ml/datasets/Car+Evaluation

2. Iris Data Set - zbiór liczący 150 elementów opisanych parametrami rzeczywistymi https://archive.ics.uci.edu/ml/datasets/iris

3. Adult Data Set - zbiór liczący 48 842 elementy opisane parametrami słownikowymi oraz całkowitoliczbowymi http://archive.ics.uci.edu/ml/datasets/Adult

Zbiory Iris i Car posiadają kolumnę określającą ich klasy co zostanie wykorzystane w przypadku testowania problemu klastrowania jako klasyfikacji. Zbiory zostały dobrane tak, aby rozwiązywać problem grupowania korzystając z różnej ilości danych oraz korzystać z różnych rodzajów parametrów (słownikowych, całkowitoliczbowych, rzeczywistych).

Wykorzystywane do grupowania algorytmy opierają się na pojęciu odległości pomiędzy punktami i jest to najczęściej odległość euklidesowa. Wymaga to, by składowe analizowanych punktów były typu numerycznego. W rzeczywistych zbiorach danych wiele parametrów jest typu słownikowego. Najprościej jest kodować typ binarny: wartość *true* jako 1 i wartość $false$ jako 0. Takie kodowanie jest dobrze określone - dla dowolnych $x, y \in \{0, 1\} : (x - y)^2 \in \{0, 1\}$ i posiada sensowną interpretację - wartość składnika $(x - y)^2$ jest równa 1 dla $x \neq y$ i równa 0 dla $x = y$. W przypadku parametrów słownikowych o liczbie możliwych wartości $n > 2$ dokonamy zamiany ich na n parametrów typu binarnego.

4 Metody przeprowadzania eksperymentów

W celu porównania jakości grupowania sieci SOM i algorytmów klasycznych, przeprowadzimy następująca procedurę:

1. Na wybranym zbiorze danych wyznaczymy klastry metodami klasycznymi: k-średnich, DBSCAN oraz najdalszego sąsiedztwa.

2. Tego samego zbioru danych użyjemy do nauczenia sieci SOM.

3. Mając wyznaczone klastry danych, możemy wyznaczyć i porównać ich jakość. Wykorzystamy dwa sposoby ewaluacji jakości klastra: wewnętrzny i zewnętrzny.

   (a) Ewaluacja wewnętrzna nie korzysta w żaden sposób z zewnętrznej wzorcowej klasyfikacji i mierzy jedynie jakość klastrowania jako podziału spełniającego następujący warunek: elementy należące do tego samego klastra powinny być do siebie nawzajem dużo bardziej podobne niż elementy należące do różnych klastrów. W celu pomiaru jakości klastrów wykorzystamy dwa powszechnie używane w tym celu wskaźniki:

      - indeksu Dunna
      - indeksu Daviesa - Bouldina

   (b) Metody zewnętrzne korzystają z zadanej - nomen omen - z zewnątrz klasyfikacji uznanej za wzorcową. Taka klasyfikacja dana jest dla dwóch spośród trzech wykorzystywanych przez nas zbiorów danych. Dla tych zbiorów, w celu wyznaczenia zgodności uzyskanej klasyfikacji ze wzorcową, wykorzystamy dwa wskaźniki powszechnie wykorzystywane do oceniania jakości testów statystycznych:

      - indeks Randa
      - F - indeks

   O ile klastry dla metod klasycznych dane są jednoznacznie, to dla sieci SOM będziemy klastrować na dwa sposoby, opisane w sekcji 2: traktując każdy neuron jako osobny klaster, lub na nauczonej sieci SOM wykonać klasyczne grupowanie.

4. Powyższe operacje powtórzymy dla wybranych przez nas zbiorów danych oraz dla różnych zestawów parametrów wykorzystywanych algorytmów:

   - liczby neuronów w sieci SOM i promienia sąsiedztwa
   - liczby klastrów w przypadku grupowania neuronów sieci SOM

- liczby klastrów k w algorytmie k-średnich
- poziomu w metodzie najdalszego sąsiedztwa
- maksymalnego promienia sąsiedztwa oraz minimalnej liczby punktów do utworzenia grupy w algorytmie DBSCAN

Tak zaplanowany eksperyment pozwala na porównanie różniących się od siebie paradygmatów grupowania za pomocą spójnych metryk oceny jakości. Powtórzenie procedury dla wybranych zbiorów danych i różnych parametrów pozwoli porównać jakość grupowania w różnych sytuacjach badawczych. Uzyskane wnioski i korelacje opiszemy szczegółowo w dokumentacji końcowej projektu.