Indukcyjne Metody Analizy Danych
laboratorium

Ćwiczenie 3. Wybrane metody klasteryzacji w oparciu o system R

opracował: P.B. Myszkowski * data aktualizacji: 09.04.2018

Cel ćwiczenia

Zapoznanie się z systemem R wspierającym statystyczne obliczenia i metody uczenia maszynowego, na przykładzie zadania uczenia nienadzorowanego, zagadnienia klasteryzacji (czasem zwaną grupowaniem).

Realizacja ćwiczenia

- Zapoznanie się z systemem R
- Zapoznanie się z następującymi algorytmami klasteryzacji: k-means i PMA
- Wybór 4 zbiorów danych do przetestowania (3 zbiory z poprzednich zadań + jeden zbiór typowo do klasteryzacji)
- Porównanie dwóch wspomnianych metod klasteryzacji ze sobą na wszystkich wybranych zbiorach
- Zbadanie jakości klasteryzacji dla 3 wybranych miar typowych dla klasteryzacji
- Analiza uzyskanych wyników
- Sporządzenie sprawozdania z ćwiczenia

Informacje pomocnicze

Do tej pory zajmowaliśmy się klasyfikacją, gdzie celem było zbudowanie modelu, który pozwoliłby na zbudowaniu modelu przypisania wektorowi danych wejściowych jednej z narzuconych klas. Teraz przechodzimy do zagadnienia klasteryzacji, gdzie klasy nie są z góry znane. Zadanie polega na pogrupowaniu danych uczących i to, jakie grupy (klastry) wyłonią się w trakcie działania algorytmu stanowi podstawowy problem.

Do przebadania są dwie metody – tworzące partycje, a nie hierarchiczne: k-means oraz PAM (*Partition Around Medoids*), przy użyciu systemu R. Celem zadania jest zapoznanie się z zestawem narzędzi do klasteryzacji, indentyfikacja parametrów algorytmów i sprawdzenie ich wpływu na uzyskane wyniki grupowania.

Przy wyborze zbiorów danych sugeruje się wybór trzech z poprzednich zadań, oraz jednego „typowego” dla klasteryzacji. Przykładowe zbiory, które docelowo przeznaczone są do zadania klasteryzacji można znaleźć pod adresem: http://archive.ics.uci.edu/ml/datasets.html?format=&task=clu&att=&area=&numAtti =&numIns=&type=&sort=nameUp&view=table

Dodatkowo, jako zbiory do badania klasteryzacji nadają się zbiory dla klasyfikatorów, też ciekawe może być porównanie znalezionych grup do ustalonych z góry klas dla klasyfikatora.
Należy pamiętać, jaka jest różnica pomiędzy zadaniem klasteryzacji i klasyfikacji. Inaczej oceniamy klasteryzację (należy wybrać 3 miary). Proszę rozpatrzyć miary DBI, Rand, Dunn czy Silhouette. Więcej na temat miar klasteryzacji w R: https://cran.r-project.org/web/packages/clusterCrit/vignettes/clusterCrit.pdf. Dla zbiorów oznaczonych (z

identyfikatorami klas) możemy ponadto zastosować miarę ***Purity***, która określa jak utworzone klastry mają się do klas.

Kroswalidacja standardowo przy klasteryzacji nie jest wykorzystywana, jednak jeśli stosujemy miary klasyfikacji (Fsc,Acc itp.) lub/i chcemy sprawdzić jak klasteryzacja ma się do klasyfikacji (np. Purity) należy rozważyć jej użycie. Warto sprawdzić wtedy jaki może mieć ona wpływ na utworzone klastry oraz ich ”jakość”.

Do oceny klasteryzacji stosuje się inne miary niż dla klasyfikacji. Należy je zidentyfikować, zrozumieć i zastosować w praktyce do uzyskanych wyników klasteryzacji. Należy wybrać minimum 3 miary, które trzeba umieć opisać (wzory!), zrozumieć w działaniu i praktyce.

Ocena ćwiczenia (max 10pkt)

2pkt	Zaznajomienie się z systemem R
2pkt	Opis działania k-means i PAM. Identyfikacja i dobór wartości parametrów.
2pkt	Porównanie działania metod klasteryzacji na 4 zbiorach (3 klasyfikacja + 1 klasteryzacja) – tabelki, wnioski. Uwaga! zbiory „weather” lub ”iris” nie stanowią dla nas wyzwania.
2pkt	Dokładna analiza zbiorów w kontekście 3 miar klasteryzacji + Purity.
2pkt	Graficzne przestawienie uzyskanych wyników. Wykresy i diagramy wskazane.

Pytania pomocnicze

1. Czy dane muszą być dyskretyzowane i/lub normalizowane?
2. Czy kroswalidacja jest potrzebna?
3. Czym różnią się oba algorytmy?
4. Który z algorytmów jest podatniejszy na szum w danych i „*outliery*”?
5. Czy istnieje potrzeba powtarzania uzyskanych wyników?
6. Czy sposób mierzenia odległości (miar) wpływa na skuteczność algorytmów?
7. Co mierzą wskazane miary jakości klasyfikacji i jakie są wartości „optymalne”. Np. jakie wartości może przyjąć miara ABC gdy mamy tylko jeden klaster, a jaką wartość jeśli mamy tyle klastrów co instancji (danych)?

Literatura

1.Cichosz P. ”Systemy uczące się”, WNT Warszawa
2.Eksploracja danych (seria wykładów) http://wazniak.mimuw.edu.pl/index.php?title=Eksploracja_danych - głównie wykłady o grupowaniu
3.Zasoby Internetu: clustering, data mining
4.http://www.project-r.org – strona projektu R
5.https://cran.r-project.org/web/packages/clusterCrit/vignettes/clusterCrit.pdf – miary klasteryzacji w projekcie R

W razie pytań, uwag czy komentarzy proszę o kontakt z prowadzącym zajęcia.
P. Myszkowski