Społeczeństwo i Edukacja
Society and Education

Wydawca / Publisher:
Instytut Studiów
Międzynarodowych
i Edukacji HUMANUM
www.humanum.org.pl

25 (2) 2017
ISSN: 1898-0171

Karolina Sala
Uniwersytet Warmińsko-Mazurski w Olsztynie, Poland
E-mail: karolinasala1@gmail.com

Przegląd technik grupowania danych i obszary zastosowań / *A review of clustering techniques and areas of application*

Abstract
The paper presents an overview of various clustering techniques used in data mining. Clustering is an unsupervised learning problem that is used to identify groups in a set of unlabeled data. Data is grouped by probability so that objects of the same group / cluster have similar properties / characteristics [1]. This article aims at exploring and comparing different clustering algorithms. Grouping is used in many areas, including machine learning, pattern recognition, image analysis, information retrieval.
Keywords: cluster analysis, hierarchical clustering, k-means.

1. WSTĘP

Klasteryzacja, zwana analizą skupień, jest procesem grupowania zbioru obiektów w taki sposób, aby obiekty powiązane w tej samej grupie były jak bardziej podobne w pewien szczególny sposób, a jak najmniej powiązane z obiektami pozostałych grup.[2] Analiza skupień odgrywa ważną rolę w wielu dziedzinach badawczych. Techniki są przydatne w różnych dziedzinach takich jak przetwarzanie obrazu czy eksploracja danych. Przykładem może być dziedzina medycyny, gdzie analiza pozwala na grupowanie chorób, symptomów oraz metod ich leczenia. Klasteryzacja odgrywa ważną rolę w aplikacjach zawierających dużą ilość informacji takich jak: aplikacje baz danych, obliczeniowe, do analizy sieci web, marketingu, typu crm.

Istnieje kilka technik klasteryzacji. Podejście każdej z nich polega na znalezieniu środka klastrów, który będzie reprezentować każdy klaster. Środek klastra reprezentuje wektor wejściowy, który może wskazywać, który klaster od niego zależy, mierząc podobieństwo między wektorem wejściowym a środkiem klastra i określenie, który klaster ma najbliższy lub najbardziej podobny[3]. Problem klasteryza-

cji nie może być rozwiązany przez jeden konkretny algorytm, ale wymaga różnych algorytmów, które odmiennie określają czym są klastry i jak skutecznie je znaleźć.

2. TECHNIKI KLASTERYZACJI

2.1. METODY HIERARCHICZNE

Klasteryzacja hierarchiczna jest metodą grupowania danych w różnej skali poprzez tworzenie drzewa klastrów zwanego dendrogramem. Drzewo jest wielopoziomową strukturą hierarchiczną, w której to klastry jednego poziomu tworzą (po połączeniu) klaster na poziomie wyższym. Dzięki temu możemy uzyskać odpowiedni do naszych potrzeb stopień klasteryzacji. Można wyróżnić dwa sposoby klasteryzacji hierarchicznej:

1. Aglomeracyjna: Zwane także podejściem od dołu. Zaczyna się od każdego obiektu tworzącego odrębną grupę. Tworzą macierz podobieństw klasyfikowanych obiektów. Łączy kolejno obiekty lub grupy, które są ze sobą podobne, dopóki wszystkie grupy nie zostaną połączone w jeden (najwyższy poziom hierarchii).
2. Deglomeracyjna: Zwane także podejściem od góry. Zaczyna od skupienia obejmującego wszystkie obiekty, a następnie w kolejnych krokach dzielą je na mniejsze i bardziej jednorodne skupienia aż do momentu, gdy każdy obiekt stanowi samodzielne skupienie [2]

Główną zaletą grupowania hierarchicznego jest brak wstępnych informacji na temat liczby wymaganych klastrów. Jest łatwa do wdrożenia i daje najlepszy rezultat w niektórych przypadkach. Wady polegają na tym, że algorytm nigdy nie może cofnąć, co prowadzi do mniejszego kosztu obliczeniowego. Jednak takie techniki nie pozwalają poprawić błędnych decyzji. Czasami trudno jest zidentyfikować poprawną liczbę klastrów przez dendrogram[4].

2.2. METODY K-ŚREDNICH

Algorytm K-średnich jest powszechnie stosowaną techniką grupowania, opartą na podziale, która używana jest do znalezienia numeru klastrów, które często reprezentowane są przez ich centroidy (środek danej grupy).

Zasada działania: Na początku wczytane dane nie mają przypisanych żadnych etykiet mogących je zidentyfikować. Następnie należy dane pogrupować. Nie posiadają oznaczeń, więc grupowanie pozwala jedynie zwizualizować ich skupiska. Obiekty zostają podzielone na 3 grupy, z racji tego, że założyliśmy, iż liczba centroidów będzie równa 3. Umieszczenie centroidów ma duże znaczenie i jest od tego uzależniony wynik działania algorytmu. Kolejny zadaniem jest ustalenie przynależności punktów do naniesionych centroidów. Aby to zrobić należy obliczyć średnie odległości punktów od centroidów. Dane znajdujące się najbliżej danego centroida są do niego przypisywane. Dla lepszego zobrazowania oznaczymy je symbolem: x. W tym kroku następuje uaktualnienie centroidów. Nowe położenie środków danych grup zostaje ustalone na podstawie średniej arytmetycznej

wszystkich punktów w niej się znajdujących. Ostatnim zadaniem algorytmu jest powtarzanie powyższych kroków, aż do osiągnięcia kryterium zbieżności.

Metodę k-średnich wykorzystuje się do analizy dużych ilości danych, a jej istota polega na zredukowaniu nadmiernej ilości nagromadzonych informacji do kilku podstawowych kategorii, co pozwala na łatwe zorientowanie się w danym zjawisku, wyciągnięcie wniosków uogólniających. Zastosowanie metody k-średnich daje możliwość ustalenia typologii w zakresie badanych obiektów oraz określenie jednorodnych przedmiotów analizy, w której łatwiej jest wyodrębnić czynniki systematyczne oraz ewentualne związki przyczynowo-skutkowe. Jej zastosowanie może prowadzić do zmniejszenia nakładów czasu i kosztów badań przez ograniczenie rozważań do najbardziej typowych faktów, zjawisk czy obiektów przy stosunkowo niewielkich stratach informacji [5].

K-średnich jest prostym algorytmem, który można zastosować do rozwiązania wielu problemów. Zaletą jest intuicyjność, łatwy do zaimplementowania. Algorytm jest bardzo wrażliwy na początkowe losowe rozmieszczenie danych oraz umiejscowienie centroidów. Czasami trzeba wielokrotnie go uruchamiać aby zredukować występowanie tego zjawiska.

2.3. METODA OPARTA NA ALGORYTMIE EM (ANG. EXPECTACTION-MAXIMIZATION)

Celem tych metod jest wykrycie skupień obserwacji (lub zmiennych) i przyporządkowanie obserwacji do skupień. Mogą być dzielone lub hierarchiczne, w zależności od struktury lub modelu, które hipotetycznie dotyczą zestawu danych i sposobu ich udoskonalenia w celu identyfikacji podziału. Są one bliższe algorytmom opartym na gęstości, ponieważ rozwijają poszczególne klastry, aby udoskonalony model uległ poprawie. Czasem jednak zaczynają się na określonej liczbie klastrów i nie używają tej samej koncepcji gęstości.

Typowym przykładem analizy tego rodzaju są badania marketingowe, gdzie dla dużej próby respondentów zbierane są pomiary pewnej liczby zmiennych opisujących zachowania konsumenckie. Celem badania jest utworzenia „segmentacji rynku," tzn. wyznaczenie grup respondentów, którzy są w jakiś sposób do siebie podobni (to znaczy podobni w obrębie tego samego skupienia) w odróżnieniu od konsumentów z innych grup.

Każdy klaster jest określony przez parametryczny rozkład prawdopodobieństwa. Obiekty są przypisywane do klastra według ich średniej wartości z jakąś masą związaną z obiektami. Rozpoczynamy od wstępnego założenia wektora parametrów, który losowo wybierany jest na podstawie wartości średniej klastrów, a następnie etap oczekiwania i stopień maksymalizacji są stosowane do dystrybucji danych.

W odróżnieniu od klasycznej implementacji *k*-średnich, algorytm *EM* może być stosowany zarówno do zmiennych ilościowych jak i jakościowych. EM jest prosty i łatwy do wdrożenia[7].

2.4. METODA OPARTA NA ALGORYTMIE DBSCAN (ANG. DENSITY BASED SPATIAL CLUSTERING OF APPLICATIONS WITH NOISE)

Najpopularniejszą metodą klastrowania opartą na gęstości jest DBSCAN. W przeciwieństwie do wielu nowszych metod charakteryzuje się dobrze zdefiniowanym modelem klastra, którym są grupy obiektów połączonych z zadaną gęstością.

Aby dobrze zrozumieć algorytm należy wprowadzić dwa nowe pojęcia:

- Eps (ε) - Jest to promień sąsiedztwa punktu określonego według przyjętej metryki.
- MinPts - Oznacza to minimalną liczbę punktów w klastrze oraz w sąsiedztwie punktów leżących wewnątrz klastra.

Główną zaletą algorytmu DBSCAN jest fakt, że nie wymaga on określenia liczby klastrów w przeciwieństwie do algorytmu k-średniej. Algorytm ten jest zdecydowanie szybszy i efektywniejszy niż standardowa klasteryzacja. Ponadto wyróżnia go możliwość znalezienia dowolnie ukształtowanych klastrów, a nie tylko sferycznych. Nie jest wyzwaniem dla algorytmu DBSCAN odnalezienia klastra całkowicie otoczonego przez inny klaster. Głównym problemem standardowej klasteryzacji jest nieprawidłowa redukcja fałszywych wskazań. DBSCAN radzi sobie z nią o wiele lepiej, gdyż jest mniej czuły na łączenie klastrów. Nieco oddalone artefakty od centrum klastra nie są do niego dołączane, co zapewnia dokładniejsze kształty i lepsze cechy kształtu i tekstury wykrytych grup.

Wadami algorytmu jest między innymi to, że nie jest on deterministyczny - jego wyniki zależą od kolejności w jakiej przeglądane są dane. Konsekwencją tego może być fakt, że jeśli przypisany jest już punkt graniczny do jednej grupy, to w późniejszym czasie może znaleźć się w innej, o ile będzie leżał dostatecznie blisko niej. Na szczęście takie sytuacje się zdarzają bardzo rzadko i nie mają dużego wpływu na efekt końcowy.

3. OBSZARY ZASTOSOWAŃ

Algorytmy grupowania mogą być stosowane w wielu dziedzinach, na przykład[6]:

1. Marketing: znajdowanie grup klientów o podobnych zainteresowaniach i zachowaniach, biorąc pod uwagę dużą bazę danych, zawierającą rejestr zakupów i aktywności;
2. Medycyna: analiza aktywności przeciwbakteryjnej, obrazowanie medyczne;
3. Finanse: prognozowanie akcji na rynku, kursy walut, ocena kredytowa;
4. Biologia: klasyfikacja roślin i zwierząt oraz ich cech, genetyka;
5. Planowanie miast: identyfikacja grup domów według ich typu, wartości, położenia geograficznego;
6. Ubezpieczenia: identyfikowanie grup z ubezpieczeniem motoryzacyjnym o wysokim średnim koszcie roszczenia, identyfikowanie oszustw;
7. Badania trzęsienia ziemi: grupowanie obserwowanych epicentrum trzęsienia ziemi w celu identyfikacji niebezpiecznych stref;

8. Informatyka: rozwój oprogramowania, rozpoznawanie obrazu;
9. WWW: klasyfikacja dokumentów, gromadzenie danych dziennika internetowego w celu wykrycia grup podobnych wzorców dostępu.

4. WNIOSKI

Podstawowym celem techniki wyszukiwania danych jest wyodrębnienie użytecznych i znaczących informacji lub wiedzy w dużych bazach danych. Ze względu na dużą ilość danych, trudno jest rozróżnić odpowiednie dane z ogromnej ilości danych. W rozwiązywaniu tego typu problemów przydatna jest metoda klasteryzacji. Podobne obiekty umieszczane są w jednym klastrze, a te, które nie są podobne, umieszcza się w innym klastrze[7]. Klastrowanie można przeprowadzić za pomocą różnych algorytmów, takich jak algorytmy oparte na hierarchii, modelu, gęstości, metod k-średnich.

LITERATURA

[1] Priyanka Sharma "Comparative Analysis of Various Clustering Algorithms" , pp.107-112, 2015

[2] https://en.wikipedia.org/wiki/Cluster_analysis

[3] Manish Verma, Mauly Srivastava, Neha Chack, Atul Kumar Diswar, Nidhi Gupta," A Comparative Study of Various Clustering Algorithms in Data Mining," International Journal of Engineering Reserch and Applications (IJERA), Vol. 2, Issue 3, pp.1379-1384, 2012.

[4] Brinda Gondaliya, "Review paper on clustering techniques", International Journal of Engineering Technology, Management and Applied Sciences, Vol 2, Issue 7, pp.234-237, 2014

[5] J. A. Hartigan and M. A. Wong (1979) „A K-Means Clustering Algorithm", Applied Statistics, 28, 100.

[6] Mamta Mor, "A Review on Various Clustering Techniques in Data Mining", International Journal of Computer Science & Communication Networks,Vol 6(3), pp.138-142, 2016

[7] Kavita Nagar, "Data Mining Clustering Methods: A Review " International Journal of Advanced Research in Computer Science and Software Engineering, Vol 5, Issue 4, pp. 575-579, 2015