Technologia

Rozwój technologii we wszystkich dziedzinach życia i wszelkie usługi przez nie dostarczane powoduje generowanie dużych ilości wielowymiarowych danych. Wszelkie formy usług dostarczanych przez sieć. Takich jak wyszukiwarki, sklepy, blogi, czy miliardy stron internetowych. Wszystkie dostarczają regularnie ogromne ilości danych.

Wraz z rozwojem technologii i usług produkujących dane, rozwijają się także środki do ich przechowywania. To nie tylko daje miejsce na przechowywanie produktów rozwijających się technologii, ale także prowokuje to do przenoszenia zbiorów dotychczas przechowywanych w bardziej „analogowy" sposób. Dokumenty, katalogi, obraz cyfrowy czy wideo.

Stosowanie elektronicznych mediów do przechowywania danych daje ogromny potencjał dla rozwoju automatycznej analizy danych, klasyfikacji i metod wyszukiwania. Na ogół jednak strumienie danych są nieuporządkowane, a fakt, że dziennie przybywa ich terabajty, bardzo to utrudnia ich analizę.

Taka ilość i różnorodność danych wymaga zaawansowanych metodologi aby automatycznie zrozumieć, przetworzyć i podsumować dane.

Wzrost ilości jak i różnorodności danych wymaga zaawansowanych metodologii aby automatycznie zrozumieć, przetworzyć i podsumować dane. Techniki analizy danych można ogólnie podzielić na dwa główne typy [z3]: (I) Badawczy lub opisowy, co oznacza, że badacz nie ma z góry określonych modeli lub hipotez, ale chce zrozumieć ogólne właściwości lub strukturę wielo-wymiarowych danych, oraz (II) potwierdzający lub wnioskujący, co oznacza, że badacz chce potwierdzenia ważności hipotezy, modelu lub zestaw założeń. Wiele technik statystycznych zostało zaproponowanych do analizy danych. Takich jak: analiza wariacji, regresji liniowej, analizy dyskryminacyjnej, korelacji kanonicznej, skalowanie wielowymiarowe, analiza czynnikowa, analza głównego składnika oraz analiza skupień.

Klasteryzacja a Klasyfikacja

Ważne jest aby zrozumieć różnicę pomiędzy klasteryzacją (klasyfikacją bez nadzoru), a analizą dyskryminacyjną (klasyfikacja nadzorowana). W klasyfikacji nadzorowanej posiadamy zbiór oznaczonych etykietami (pre-sklasfikowanych) rekordów. Problem polega na oznaczeniu nowo napotkanych i nie oznaczonych rekordów odpowiednimi etykietami. Na ogół, dane, oznaczone etykietami (treningowe), wektory są używane do nauki opisów klas, które później są używane do etykietowania nowych rekordów. Natomiast klasteryzacja polega na pogrupowaniu kolekcji nieoznaczonych wektórów w klastry posiadające jakieś znaczenie. Tutaj także są etykiety, ale wywodzą się one czysto z danych, które grupowano.

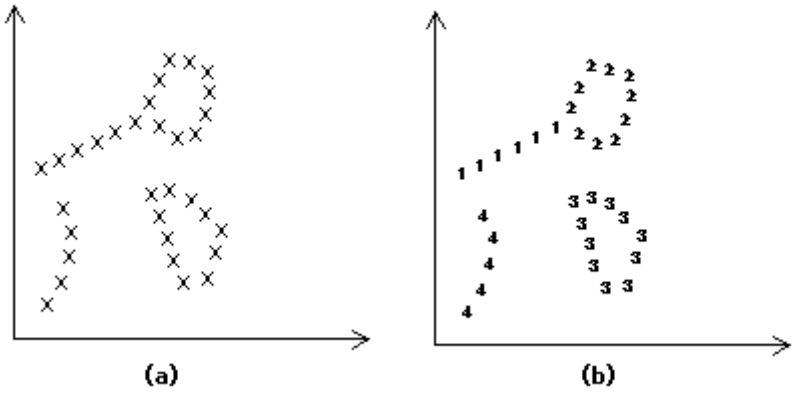
Klasteryzacja jest przydatna w kilku sytuacjach. Taich jak: odkrywcza analiza wzorów, grupowanie, podejmowanie decyzji, uczenie maszynowe. W tym data-mining, wyszukiwanie dokumentów, segmentacja obrazu oraz klasyfikacja rekordów. Na ogół w takich sytuacjach, niewiele jest informacji wstępnych (np. modeli statystycznych) na temat dostępnych danych. To w ramach tych ograniczeń, tworzenie klastrów jest szczególnie odpowiednią metodologią badań współzależności między punktami danych aby dokonać oceny (być może wstępnej) ich struktury.

Klasteryzacja – jak to działa

Celem grupowania (klasteryzacji) danych, znanego także jako analizy skupień, jest odkrycie naturalnych ugrupowań zbioru rekordów, punktów lub obiektów.

Webster [z4] definiuje analiza skupień jako „Statystyczna technika klasyfikacji dla odkrywania czy jednostki populacji przynależą do różnych grup, poprzez porównania ilościowe wielu charakterystyk."

Przykład klasteryzacji przedstawiono na rysunku 1. Celem jest opracowanie automatycznego algorytmu, który odkryje naturalne ugrupowania (rys 1b) w nie oznaczonych danych (rys 1a). Operacyjną definicję klastrów można sformułować następująco [z3]: Mając n obiektów, znajdź K grup w oparciu o miarę podobieństwa, tak aby podobieństwo pomiędzy obiektami tej samej grupy było wysokie, natomiast niskie dla obiektów z różnych grup. Lecz czym jest miara podobieństwa ? Jaka jest definicja klastra ? Z rys 1 wynika, że klastry mogą różnić się pod względem kształtu, wielkości oraz gęstości. Obecność szumu w danych czyni wykrywanie klastrów jeszcze trudniejszym. Idealny klaster może zostać zdefiniowany jako zbiór punktów, zwarty i odizolowany. Jednakże w rzeczywistości, klaster to subiektywny byt będący w oku patrzącego, którego znaczenie i interpretacja wymaga wiedzy na temat domeny. Gdy jednak ludzie są doskonali w wyszukiwaniu klastrów w dwóch i prawdopodobnie w trzech wymiarach, potrzebujemy zautomatyzowanych algorytmów dla danych wielowymiarowych. To i niewiadoma ilość klastrów stały się wyzwaniem, które wynikło tysiącami algorytmów, które zostały opublikowane i nadal są.



Rysunek 1

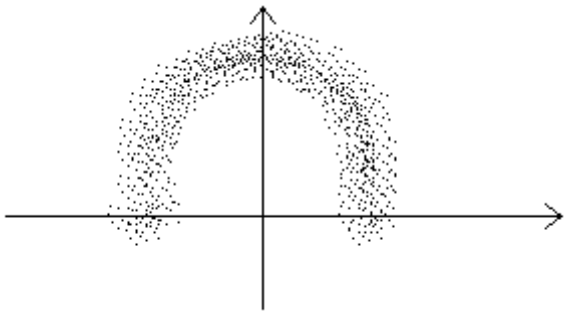
Następujące notacje używane są w dalszej treści dokumentu:

- *Rekord* (lub *wektor cech*, *obiekt* lub *punkt*) **x** jest pojedynczym elementem danych, użytym w algorytmie klasteryzacji. Zwykle jest wektorem d miar: **x** = (x_1, …, x_d).
- Indywidualne komponenty x_i rekordu **x** zwane są *cechami* (lub *atrybutami*).
- *d* to *wymiarowość* rekordu lub przestrzeni rekordów.
- *Zbiór rekordów* oznaczony jako $\mathcal{X}$ = {$\mathbf{x}_1$, …, $\mathbf{x}_n$}. W wielu przypadkach, zbiór który ma podlegać klasteryzacji jest postrzegany jako macierz $n \times d$ rekordów.
- *Klasa*
- *Twarde* techniki klasteryzacji każdemu rekordowi x_i przypisują etykietę klasy l_i Identyfikując jego klasę. Zbiorem wszystkich etykiet dla zbioru rekordów $\mathcal{X}$ jest $\mathcal{L}$ = {l_i, …, l_n}, z l_i {1, …, k} $l_i \in \{1, \ldots, k\}$, gdzie *k* jest liczbą klastrów.
- Procedury *rozmytej* klasteryzacji przypisują każdemu rekordowi stopień przynależności f_{ij} do danego klastra wyjściowego *j*

- *Miękkie* techniki klasteryzacji przypisują każdemu rekordowi $\mathbf{x}_i$ stan przynależności l_{ij} do danego j-tego klastra. Przy czym $l_{ij} \in \{0, 1\}$, gdzie 0- nie należy, 1-należy do klastra.
- *Miara odległości* jest miarą określającą podobieństwo rekordów w przestrzeni ich atrybutów

Reprezentacja rekordów, wybór i ekstrakcja cech

Brak konkretnych przewodników na temat poprawnej reprezentacji rekordów i doboru cech użytych dla specyficznych sytuacji. Na ogół proces generowania rekordów nie podlega bezpośredniej kontroli. Rolą użytkownika jest gromadzenie faktów i przypuszczeń, ewentualnie dokonać wyboru i ekstrakcji cech. Na pewno dokładna analiza dostępnych cech i dostępnych przekształceń (nawet tych prostych) może dać znacznie lepsze wyniki. Może to zadecydować czy otrzymamy proste i dobrze zrozumiałe klastry, czy skomplikowaną strukturę, której prawdziwą naturę trudno odgadnąć.



Rysunek 2 ukazuje prosty przykład. Punkty, w tej przestrzeni 2D cech, są zorganizowane w kształt mniej więcej równo oddalony od jednego punktu środka. Jeśli wybrać współrzędne Kartezjańskie, większość algorytmów klastrowania prawdopodobnie podzieli tą figurę na dwa lub więcej klastrów. Jeśli by jedna użyć reprezentacji koordynatów polarnych, uzyskanie pojedynczego klastra ma większe prawdopodobieństwo.

Rekord może mierzyć tak fizyczny obiekt (np. krzesło) jak i abstrakcyjny (np. styl pisania). Jak wspomniano powyżej, każdy obiekt reprezentowany jest jako wektor, gdzie każdy wymiar jest pojedynczą cechą. Cechy te można podzielić na ilościowe i jakościowe. Dla przykładu, jeśli waga i kolor były by dwiema użytymi cechami, wtedy (20, black) jest reprezentacją czarnego obiektu o wadze 20 jednostek. Cechy można podzielić na następujące typy.

(1) Cechy ilościowe
   A. Wartości ciągłe (np. waga)
   B. Wartości dyskretne (np. liczba komputerów)
   C. Wartości przedziału (np. czas przebiegu wydarzenia)

(2) Cechy jakościowe
   A. Nominalne lub nieposortowane (np. kolor)
   B. Porządkowy (np. ranga wojskowa, temperatura („gorąco”, „zimno”), czy intensywność głośności („ciho”, „głośno”)

Można także używać cech ustrukturyzowanych, reprezentowanych przez drzewo, gdzie rodzic jest generalizacją swoich dzieci. Dla przykładu, węzeł rodzic może być generalizacją dla dzieci oznaczonych jako „samochódy”, „autobusy”, „ciężarówki” i „motory”. Dalej, węzeł „samochody” może być generalizacją samochodów typu „Toyota”, „Ford”, „Mercedes”

Ważne jest aby wyizolować tylko te najbardziej wartościowe, opisowe i dyskryminujące cechy z zestawu wejściowego, a następnie indywidualnie poddać kolejnym przetwarzaniom analizy. Techniki wyboru cech wyznaczają podzbiory istniejących cech. Natomiast techniki ekstrakcji obliczają nowe cechy z już istniejących. W obu przypadkach celem jest polepszenie jakości klasyfikacji i/lub efektywności obliczeniowej. Dobór cech to bardzo obszerny temat, jednakże w klasteryzacji często zmuszeni jesteśmy na dokonywanie wyborów „na wyczucie" a nie rzadko jest to proces prób i błędów, gdzie wybierane są różne zestawy cech obiektów i wybór jest oceniany po zakończeniu procesu klasteryzacji.

Miara podobieństwa

Jako, że podobieństwo jest podstawą definicji klastra, pomiar podobieństwa dwóch rekordów z tej samej przestrzeni cech, jest bardzo ważne dla większości procedur klasteryzacji. Z powodu różnorodności typów wartości cech i ich scalowania, miara (lub miary) podobieństwa, muszą być dobierane bardzo ostrożnie. Najbardziej powszechną praktyką jest obliczanie *miary niepodobieństwa* pomiędzy dwoma rekordami używając miary odległości zdefiniowanej na przestrzeni cech. Najczęściej w przypadku cech o wartościach ciągłych stosowana jest odległość Euklidesowa.

$$d(\mathbf{x}_i, \mathbf{x}_j) = \left(\sum_{k=1}^{d} (\mathbf{x}_{i,k} - \mathbf{x}_{j,k})^2 \right)^{1/2} = \|\mathbf{x}_i - \mathbf{x}_j\|_2$$

Które jest specjalnym przypadkiem (p=2) metryki Minkowskiego.

$$d(\mathbf{x}_i, \mathbf{x}_j) = \left(\sum_{k=1}^{d} (\mathbf{x}_{i,k} - \mathbf{x}_{j,k})^p \right)^{1/p} = \|\mathbf{x}_i - \mathbf{x}_j\|_p$$

Odległość Euklidesowa jest intuicyjnie używane do obliczeń odległości obiektów w dwu i trzy wymiarowych przestrzeniach. Sprawuje się doskonale tak z wyizolowanymi jak i złożonymi klastrami. Wadą używania tej miary jest fakt, że szybko cechy o mniejszej skali stracą na wartości zdominowane przez cechy o większej skali. Rozwiązaniem tego problemu jest normalizacja używanych wartości lub stosowanie wag.

Obliczanie odległości pomiędzy rekordami, których część lub wszystkie cechy nie są wartościami ciągłymi, staje się o wiele trudniejsze. W ekstremalnych przypadkach wartość odległości jest określana binarnie. I tutaj jednak istnieje wiele prac, które starają się rozwiązać ten problem.

Rekordy mogą być także reprezentowane używając stringów czy struktur drzewiastych. Wymaga to wtedy syntaktycznego lub statystycznego podejścia do miary odległości. Ale są one zdecydowanie mniej przydatne.

Są także miary odległości, który pod uwagę biorą otoczenie lub sąsiednie punkty zwanych kontekstem. Miarę tej odległości można określić wzorem:

$$MND(\mathbf{x}_i, \mathbf{x}_j) = NN(\mathbf{x}_i, \mathbf{x}_j) + NN(\mathbf{x}_j, \mathbf{x}_i)$$

Gdzie NN($\mathbf{x}_i$, $\mathbf{x}_j$) jest ilością sąsiadów punktu $\mathbf{x}_i$ w odniesieniu do punktu $\mathbf{x}_j$. Tzn. funkcja ta określa, którym z kolei najbliższym sąsiadem $\mathbf{x}_i$ jest $\mathbf{x}_j$. Oznacza to, że im więcej obiektów pomiędzy dwoma punktami, tym większa odległość wynikowa. Mimo, że nie miara ta nie

spełnia zasad nierówności trójkątnej, została z powodzeniem zastosowana w kilku pracach i pokazuje, że *niepodobieństwo* nie musi być miarowe.

Techinki Klastrowania.

Według [z5] mamy cztery główne techniki klasteryzacji: partycjonowanie, hierarchiczne, gęstościowe, i sieć(grid).

- Technika partycyjna organizuje dane w k klastry tak, że ich odległości do punktu reprezentującego dany klaster (k-means) jest najmniejsza. Mimo, że jest to jedna z najpopularniejszych technik: (a) nie nadają się do wykrwania klastrów nie wypukłych, (b) są wrażliwe na odstające punkty i początkową selekcję centrów klastrów, (c) z góry wymaga określenia ilości klastrów, (d) nie radzi sobie z klastrami o różnych rozmiarach i gęstościach, (e) generuje kiepskiej jakości deskryptory klastrów.
- Techniki hierarchiczne produkują zagnieżdżoną sekwencję klastrów. Hierarchia może być formowana „z góry" lub „z dołu". Każda iteracja polega na łączeniu lub rozdzielaniu klastrów na podstawie odpowiednich miar podobieństwa pomiędzy klastrami. Główną wadą tej techniki jest niemożność korekcji wyników powykonaniu operacji (dzielenie lub łączenie) oraz wysoka skomplikowalność obliczeń. Można tu wymienić takie algorytmy jak BIRCH, CURE czy CHAMELEON
- Techniki Gęstościowe grupują sąsiednie punkty w klastry bazując na lokalnej gęstości zamiast odległości pomiędzy punktami. Popularniejsze algorytmy to: DBSCAN, OPTICS, DENCLUE
- Techniki sieci (grid) organizuje przestrzeń danych w wielowymiarową sieć i agreguje punkty w komórki. Podstawowy pomysł polega na tym aby odrzucać komórki o małej gęstości, a następnie łączyć komórki o wysokiej gęstości i formować klastry. Metoda ta jest znacząco szybka dla nisko do średnio wymiarowych danych i nieźle radzą sobie z wykrywaniem klastrów o niestandardowych kształtach. Jednakże, jako że ilość komórek rośnie gwałtownie wraz z kolejnymi wymiarami, staje się to trudne i kosztowne dla wielowymiarowych danych. Znaczącymi metodami są tu: CLIQUE, MAFIA, WaveCluster.

Algorytmy Ewolucyjne w klasteryzacji

Jeśli spojrzeć na problem klasteryzacji jako na problem optymalizacji. Czyli przy zastosowaniu dowolnej z powyższych technik, chcemy uzyskać jak najlepsze wyniki, a nierzadko niektóre z nich bardzo są wrażliwe na startowe parametry. Takie jest główne założenie Algorytmów Ewolucyjnych w zadaniu klasteryzacji. Z taką drobną różnicą, że techniki zostały zredukowane do prostych funkcji oceny.

Najprostszy podział EA można wprowadzić na podstawie ostatecznych wyników jakie dostarczają na wyjściu [z6][…]. Są to klastry *twarde*, *miękkie* oraz *rozmyte*. Działają one dokładnie w ten sam sposób jak opisano wcześniej w tym dokumencie.

Drugi podział to algorytmy o *stałej* liczbie klastrów oraz *dynamicznej*. W pierwszym wypadku liczba klastrów jest z góry określona i nie zmienia się przez cały proces wykonywania algorytmu. Wartość tą użytkownik musi ustalić sam używając wiedzy o domenie w której dokonuje analizy, lub metodą prób i błędów przeprowadzając kilka analiz na tym samym zbiorze danych. Dynamiczne algorytmy natomiast starają się same określić ilość klastrów i to również podlega procesowi optymalizacji.

Kolejną dziedziną podziału algorytmów Ewolucyjnych w klasteryzacji jest typ zastosowanego chromosomu do opisu pojedynczego rozwiązania[z6] :.

- Kodowanie binarne – jest to ciąg n bitów, gdzie n to ilość rekordów podlegających analizie. W ciągu tym kolejne bity przypisane są do kolejnych rekordów, a bit pozytywny oznacza, że ten rekord jest prototypem klastra.
- Kodowanie całkowitymi – Dzieli się na dwa możliwe zastosowania. Pierwszy to ciąg n liczb, każda przypisana kolejnemu obiektowi zbioru danych zawiera numer z przedziału {1, …, k} opisujący do jakiego klastra dany obiekt przynależy. Ten sposób zapisu niesie ze sobą niebezpieczeństwo powtarzalności, ponieważ pojedyncze rozwiązanie może być zapisane na wiele sposobów. Innym sposobem na zastosowanie zapisu w postaci licz całkowitych to użycie wektora k-elementowego, gdzie każdy opisuje numer obiektu ze zbioru danych użytego jako medoidów.
- Kodowanie rzeczywistymi – w postaci tej, kolejne wartości to współrzędne prototypów w metodologii bazującej na centroidach. W takim wypadku, genotyp będzie miał długość k*d, gdzie k to liczba klastrów, a d to liczba wymiarów.
- Kodowanie hybrydowe lub niestandardowe -

Funkcje oceny.

W praktyce, każde relatywne kryterium oceniające poprawność klastra, nie będące monotoniczne względem liczby klastrów, może zostać użyte jako funkcja oceny. Jednakże dobór tej części algorytmu jest bardzo ważny jako, że kreuje on przebieg procesu jak i ostatecz wynik działania. Samych funkcji jest wiele i wymienie tutaj tylko ich mały podzbiór.

- Funkcja gęstości – średnia odległość obiektów do przypisanych im centroidów.
- Funkcja rozłączności – czyli odległość pomiędzy centroidami
- Funkcja łączności –funkcja ta sprawdza izolację klastrów średnią liczbę sąsiadów, którzy należą do tej samej grupy.

Wielokryterialność

W przeciwieństwie do uczenia się nadzorowanego, jak np. klasyfikacja, klasteryzacja, jako metoda nie nadzorowana, nie posiada „bazowej prawdy”, mówiącej czy rozwiązanie jest poprawne czy optymalne. To sugeruje, aby jakość rozwiązania była określana przez różnorodny zestaw kryteriów, zamiast pojedynczy. Aby likwidować złe wyniki w dowolnym kryterium poprawności.

Aby móc z powodzeniem zastosować wielokryterialność w algorytmie klasteryzacji, kryteria muszą zostać dobrane z pewną drobiazgowością. Rozpatrując różne aspekty jakości klasteryzacji. Częstą praktyką jest definiowanie kryteriów wręcz zaprzeczających się nawzajem. Wtedy tworzy się pojedynczą regułę, gdzie każde kryterium posiada swoją wagę ważności. Wadą tu jest fakt, że samo obliczenie stosowanych wag, może wymagać dodatkowych przebiegów aby znaleźć najlepsze rozwiązanie.

W idealnym algorytmie wielo-kryterialnym stosuje się dominację Pareto. Polega to, w skrócie, na odnajdywaniu członków populacji, nie zdominowanych przez nikogo innego. Tzw. „Pareto front”. Dany członek populacji dominuje, jeśli we wszystkich kryteriach jest równy lub w co najmniej jednym lepszy.