Moduł 3
Eksploracja danych

iBigWorld:
Innovations for Big Data in a Real World

TSNUK team

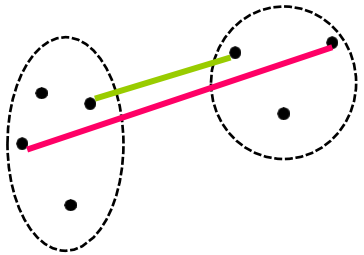
iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Pomiar zbliżeniowy

Odległość euklidesowa jest najczęstszym rodzajem definicji odległości. Jest to odległość geometryczna między punktami w przestrzeni wielowymiarowej:

$$\rho_{ij} = \left[\sum_k (x_{ik} - x_{jk})^2 \right]^{1/2}$$



gdzie: X_i , X_j są współrzędnymi i^{th} и j^{th} obiektów w k-wymiarowejprzestrzeni;

x_{il} - x_{jl} jest wartością i^{th} składnika i^{th} (or j^{th}) obiektu (i = 1, 2, ..., k; i, j = 1, 2, ..., n).

Kwadratowa odległość euklidesowa służy do przypisywania większej wagi obiektom, które są bardziej od siebie oddalone:

$$\rho_{ij} = \left[\sum_k (x_{ik} - x_{jk})^2 \right]$$

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Pomiar zbliżeniowy

Odległość miejska porównana została do odległości euklidesowej, wpływ oddzielnych dużych różnic (odstających) jest zmniejszony, ponieważ nie są one podniesione do kwadratu:

$$\rho_{ij} = \sum_k | x_{ik} - x_{jk} |$$

where: X_i , X_j są współrzędnymi i^{th} и j^{th} obiektów w k-wymiarowej przestrzeni;

x_{il} - x_{jl} wartością i^{th} składnika i^{th} (or j^{th}) obiektu (i = 1, 2, ..., k; i, j = 1, 2, ..., n).

Odległość Minkowskiego :

$$\rho_{ij} = \left[\sum_k \left| x_{ik} - x_{jk} \right|^p \right]$$

If P = 1 – odległość miejska,

if P = 2 –odległość

Euklidesowa.

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Przygotowanie danych

Przygotowanie danych to proces czyszczenia, strukturyzacji i wzbogacania surowych danych aby uzyskać żądany zbiór do analizy.

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Wyodrębnianie informacji

Wyodrębnianie informacji to proces tworzenia bazy danych zawierającej informacje wydobyte z zestawu dokumentów przygotowanych w języku naturalnym.

Typowe zadania ekstrakcji informacji:

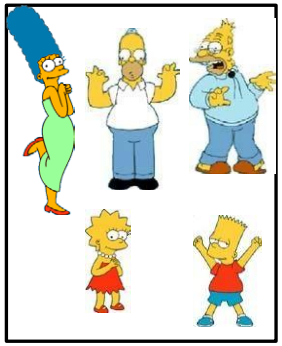
- Rozpoznawanie nazwanych jednostek — rozpoznawanie nazw jednostek (w przypadku osób lub organizacji mogą to być nazwy produktów, nazwy lokalizacji, wyrażenia czasowe i niektóre typy wyrażeń liczbowych).
- Referencje - identyfikacja wzorców fraz rzeczownikowych odnoszących się do tego samego przedmiotu.
- Wyodrębnianie terminologii - znajdowanie odpowiednich terminów dla danego wzorca.
- Wydobycie opinii lub wydobycie sentymentów – określenie tonacji tekstu (pozytywnej lub negatywnej) w procesie opisywania osoby, produktu lub usługi.

iBigWorld:2020-1-PL01-KA203-082197

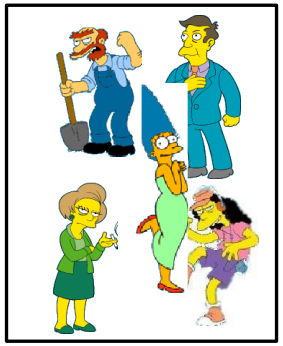
Co-funded by the Erasmus+ Programme of the European Union

Analiza skupień

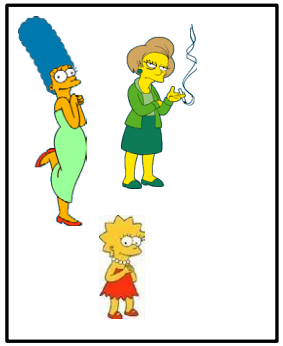
Grupowanie to dekompozycja elementów ustawionych w grupy na podstawie ich podobieństwa. Zadaniem klastrowania jest rozbicie obiektów z X na kilka podzbiorów (klastrów), w których obiekty są do siebie bardziej podobne niż obiekty z innych klastrów.



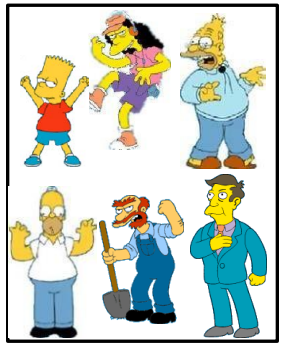
Rodzina



Współpracownicy



Kobiety



Mężczyźni

Należy zdefiniować etykiety grup

Grupowanie jest procesem subiektywnym i zależy od celu użytkownika

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Analiza klastrowa

Procedura klastrowania zależy od stopnia podobieństwa lub braku podobieństwa. Takie miary można opisać w postaci funkcji odległości, wyrażonych w postaci takiej lub innej funkcji.

Określenie podobieństwa jest właśnie uczeniem maszynowym.

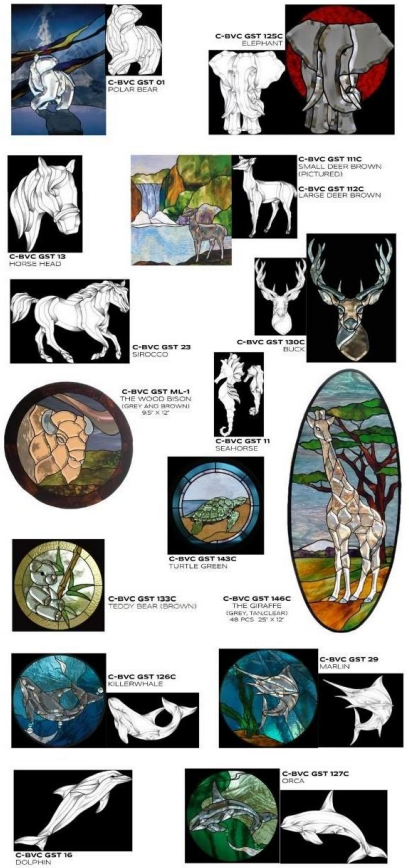
"We know it when we see it."

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Analiza skupień - zastosowanie

1. Statystyki.
2. Rozpoznawanie wzorców.
3. Uczenie maszynowe.
4. Matematyka finansowa.
5. Automatyczna klasyfikacja w różnych dziedzinach nauki, np. w archeologii, biologii (grupowanie gatunków zwierząt i roślin).
6. Marketing. Marketerzy identyfikują grupy w celu optymalizacji działań reklamowych, optymalizacji działań logistycznych.
7. Badania właściwości DNA.
8. Ubezpieczenia (w celu identyfikacji grup ludności i korelacji grup z położeniem geograficznym, zarobkami, stanem cywilnym itp.).
9. Planowanie urbanistyczne.
10. Planowanie finansowe.



iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the
Erasmus+ Programme
of the European Union

Analiza skupień - kierunki

1. Partycjonowanie w klastry zakłada dekompozycję obiektów na klastry na raz, a jeden obiekt należy tylko do jednego klastra.
Typowe metody: k-średnie, k-medoidy, CLARANS

2. Metoda klastrowania rozmytego pozwala na rozbicie istniejącego zbioru obiektów p na określoną liczbę zbiorów rozmytych, co oznacza, że ten sam obiekt może należeć do różnych klas. Przynależność charakteryzuje się stopniem przynależności, takim jak prawdopodobieństwo.
Typowe metody: k-średnie.

3. Klasteryzacja hierarchiczna dopuszcza obecność podgrup. Przebiega w kilku etapach, w wyniku czego tworzy hierarchiczne drzewo (dendrogram).
Typowe metody: Hierarchiczna, Diana, Agnes, BIRCH, ROCK, CAMELEON.

4. Podejście oparte na modelu gęstości opiera się na funkcjach łączności i gęstości.
Typowe metody: DBSACN, OPTICS, DenClue.

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the
Erasmus+ Programme
of the European Union

Analiza skupień - dane

Klastrowanie może pomóc odpowiedzieć na takie pytania jak:

1. Czy grupy klastrów (firm) będą odzwierciedlać parametry, które nie zostały uwzględnione w analizie skupień.

2.Based on the analysis, it is possible to formulate correlating rules between companies, products and consumers, and apply these rules to companies that were not included in the analysis.

Member State	Population in millions	Population % of EU	Area km^2	Area % of EU	Pop. density People/km^2
Austria	8.3	1.7%	83,858	1.9%	99
Belgium	10.5	2.1%	30,510	0.7%	344
Bulgaria	7.7	1.6%	110,912	2.5%	70
Croatia	4.3	0.9%	56,594	1.3%	75.8
Cyprus	0.8	0.2%	9,250	0.2%	84
Czech Republic	10.3	2.1%	78,866	1.8%	131
Denmark	5.4	1.1%	43,094	1.0%	126
Estonia	1.3	0.3%	45,226	1.0%	29
European Union	494.8	100%	4,422,773	100%	112
Finland	5.3	1.1%	337,030	7.6%	16
France	65.03	13.%	643,548	14.6%	111
Germany	80.4	16.6%	357,021	8.1%	225
Greece	11.1	2.2%	131,940	3.0%	84
Hungary	10.1	2.0%	93,030	2.1%	108
Ireland	4.6	0.9%	70,280	1.6%	60
Italy	58.8	11.9%	301,320	6.8%	195
Latvia	2.3	0.5%	64,589	1.5%	35
Lithuania	3.4	0.7%	65,200	1.5%	45
Luxembourg	0.5	0.1%	2,586	0.1%	181
Malta	0.5	0.1%	316	0.0007%	1,261
Netherlands	17	3.3%	41,526	0.9%	394
Poland	38.1	7.7%	312,685	7.1%	122
Portugal	10.6	2.1%	92,931	2.1%	114
Romania	21.6	4.4%	238,391	5.4%	91
Slovakia	5.4	1.1%	48,845	1.1%	111
Slovenia	2.0	0.4%	20,253	0.5%	99
Spain	44.7	9.0%	504,782	11.4%	87
Sweden	10	1.8%	449,964	10.2%	20

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Analiza skupień - dane

Dane tekstowe. Dokumenty są reprezentowane w modelu wektorowym, co oznacza, że zbiór dokumentów jest przedstawiony w postaci macierzy termin-dokument. Linie będą reprezentować poszczególne dokumenty (teksty), a kolumny będą reprezentować przykładowy zestaw danych, który zawiera wszystkie słowa z kolekcji dokumentów. Dodatkowo prowadzi się preprocessing, który polega na usuwaniu tagów html, lematyzacji i usuwaniu słów „stop”.

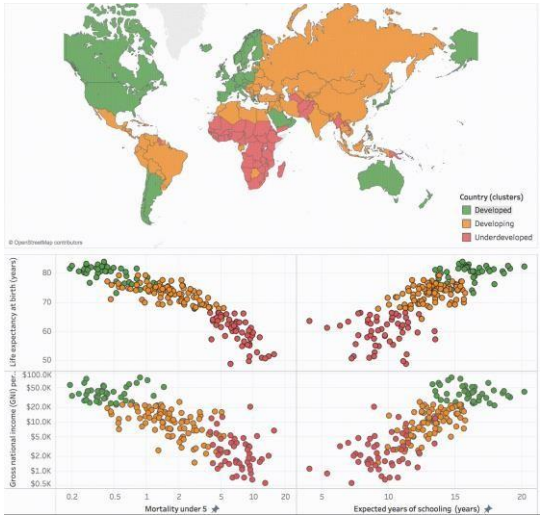
	Jane	***likes***	***coffee***	***and***	***tea***	***also***	***cookies***
1st text	**1**	**1**	**1**	**1**	**1**	**0**	**0**
2nd text	**1**	**1**	**0**	**0**	**0**	**1**	**1**

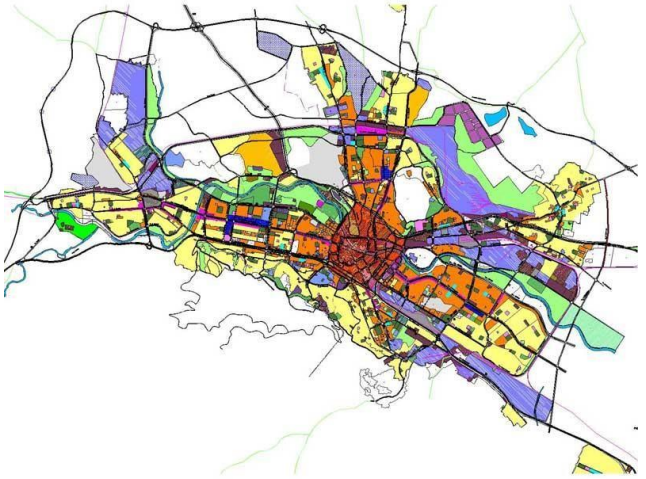
Dane tekstowe jako zbiór wektorów binarnych :d_1 = [1,1,1,1,1,0,0] and d_2 = [1,1,0,0,0,1,1].

Co-funded by the Erasmus+ Programme of the European Union

Analiza skupień - dane

Dane mapy. Dane przestrzenne (na przykład foto-, geomapy) mogą być reprezentowana jako zbiór punktów.
Takie dane często występują przy rozwiązywaniu problemu rozpoznawania wzorców.





iBigWorld:2020-1-PL01-KA203-082197

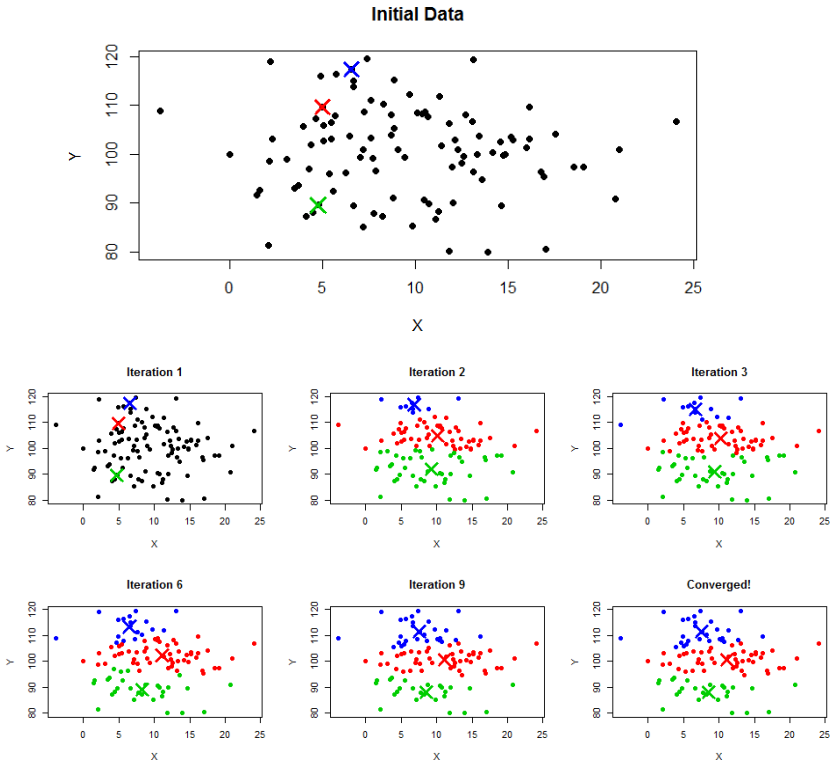
Co-funded by the
Erasmus+ Programme
of the European Union

Analiza skupień: Algorytm K-średnich

Podstawowa idea grupowania jest następująca: Mamy dany zbiór obiektów.

1. Wybierz punkty początkowe dla klastrów.
2. Połącz najbliższe punkty z centrami klastrów.
3. Przelicz ponownie centra klastrów przy założeniu, że do klastra zostały dodane nowe obiekty.
4. Po znalezieniu centrów skupień ponownie rozmieść najbliższe punkty w skupieniach.

Środek skupienia rozumiany jest jako średnia arytmetyczna skupień.



iBigWorld:2020-1-PL01-KA203-082197

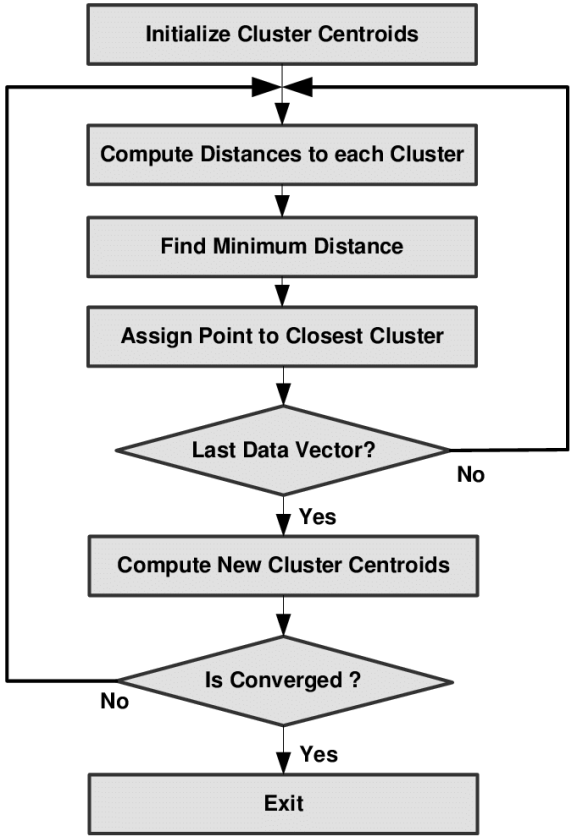
Co-funded by the Erasmus+ Programme of the European Union

Analiza skupień: Algorytm K-średnich

1. Nie ma gwarancji, że zostanie osiągnięte globalne minimum całkowitego odchylenia kwadratowego V, a nie tylko jedno z minimów lokalnych.
2. Wynik zależy od wyboru początkowych ośrodków klastrowych; jego optymalny dobór jest nieznany.
3. Liczba klastrów musi być znana z góry.

Jak można przezwyciężyć te problemy?

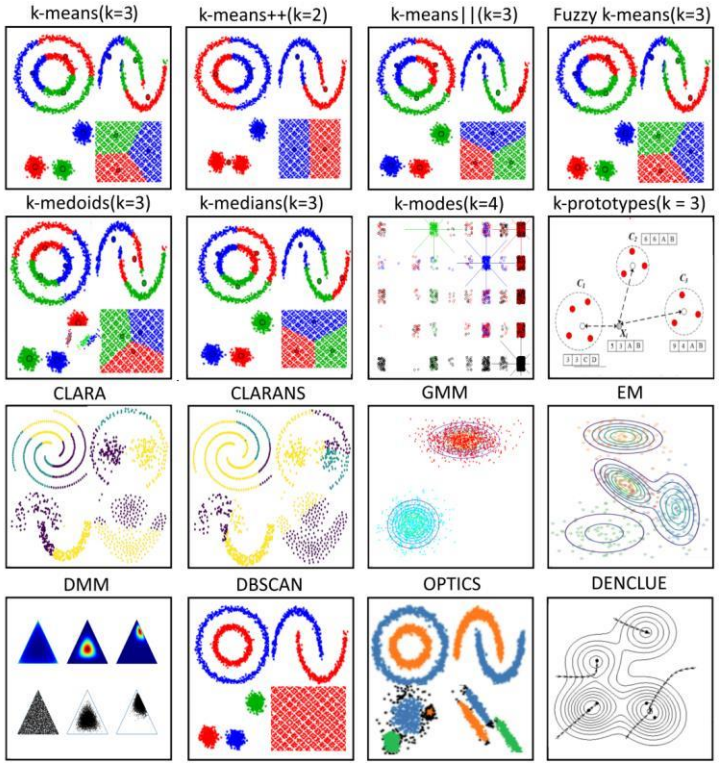
1. Uruchom algorytm wiele razy (z różnymi centrami skupień), a następnie wybierz wynik z minimalnym błędem.
2. Użyj dodatkowych modeli do oszacowania liczby klastrów.



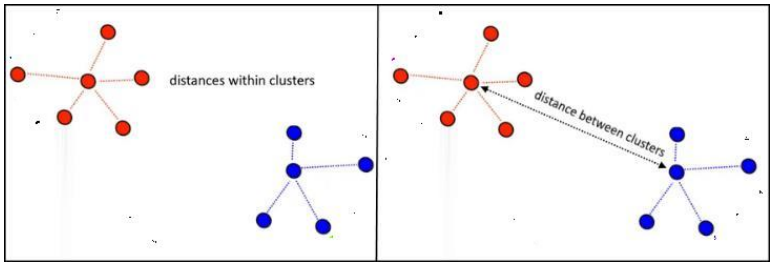
iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Analiza skupień: Typy klastrów



Różne typy klastrów prowadzą do problemu doboru optymalnej liczby klastrów oraz problemu doboru wymaganego minimum.



iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Analiza skupień: Kryterium zakończenia obliczeń

Zadanie grupowania można rozwiązać jako problem optymalizacji dyskretnej: konieczne jest przyporządkowanie liczb skupień yi do obiektów xi w taki sposób, aby wartość wybranego funkcjonału jakości przybrała najlepszą wartość. Istnieje wiele odmian funkcjonałów jakości klastrowania, ale nie ma „najbardziej poprawnej" funkcjonalności.

Średnia odległość między klastrami musi być jak najmniejsza:

$$F_0 = \frac{\sum_{i<j}[y_i = y_j]\rho(x_i, x_j)}{\sum_{i<j}[y_i = y_j]} \to min.$$

Średnia odległość między skupiskami musi być jak największa:

$$F_1 = \frac{\sum_{i<j}[y_i \neq y_j]\rho(x_i, x_j)}{\sum_{i<j}[y_i \neq y_j]} \to max.$$

Co-funded by the Erasmus+ Programme of the European Union

Analiza skupień: Kryterium zakończenia obliczeń

Jeżeli algorytm grupowania oblicza środki skupień μy, y ∈ Y, to można zdefiniować funkcjonały, które są bardziej wydajne obliczeniowo.

Suma średnich odległości wewnątrz klastra musi być tak mała, jak to możliwe:

$$\Phi_0 = \sum_{y \in Y} \frac{1}{|K_y|} \sum_{i:y_i=y} \rho^2(x_i, \mu_y) \to min.$$

gdzie $K_y = \{x_i \in X_\ell \mid y_i = y\}$ jest numerem klastra y. W tym wyrażeniu można zastosować same odległości, a nie ich kwadraty. Jeśli jednak ρ jest metryką euklidesową, to suma wewnętrzna w Φ0 nabiera znaczenia fizycznego momentu bezwładności skupienia K_y w stosunku do jego środka masy, jeśli skupienie uważane jest za ciało materialne składające się z $|K_y|$ punktów o tej samej masie.

Suma odległości między klastrami musi być jak największa:

$$\Phi_1 = \sum_{y \in Y} \rho^2(\mu_y, \mu) \to max.$$

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Analiza skupień:
Sugar & James Algorithm

W celu rozwiązania problemu kryterium zatrzymania obliczeń istnieje metoda nieparametryczna zaproponowana przez panów Sugar i James, która pozwala na takie przekształcenie funkcji jakości, aby przegięcie lub skok stały się wyraźnie widoczne. Metoda opiera się na wykorzystaniu pojęcia zniekształceń, które są oszacowaniami wariancji w ramach klasy (zgrupowania).

Podczas realizacji tej metody obliczana jest funkcja specjalna (funkcja zniekształceń), która jest wyznaczana na podstawie trzech parametrów:

1. Zniekształcenie dla danego rozwiązania klastrowego.
2. Liczba klastrów dla danego rozwiązania klastrowego.
3. Współczynnik konwersji. Zniekształcenie jest przekształcane w zniekształcenie przekształcone na podstawie współczynnika konwersji. Na koniec przeprowadzana jest analiza zachowania przekształconej funkcji zniekształcenia w zależności od liczby klastrów. Na podstawie analizy wyciąga się wniosek, które rozwiązanie klastra jest najlepsze.

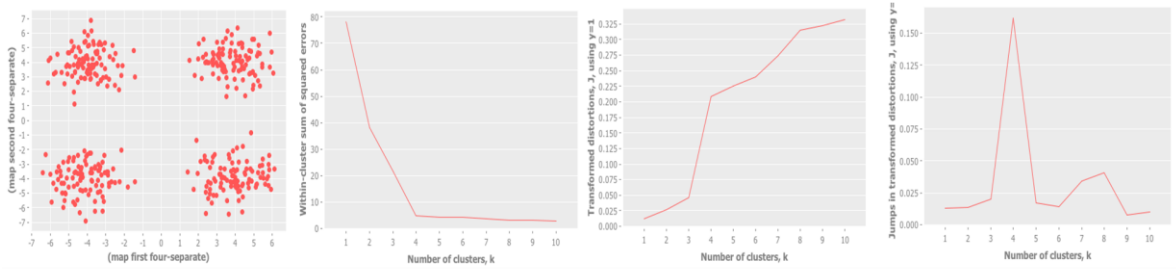
iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the
Erasmus+ Programme
of the European Union

Analiza skupień:
Sugar & James Algorithm

1. Wyznaczanie minimalnych zniekształceń. Jako minimalne zniekształcenie przyjmuje się minimalną wartość wariancji wewnątrzklastrowej, która występuje w danym rozwiązaniu klastra. Oznacza to, co następuje: dla danego rozwiązania skupień (na przykład dla 5 skupień) obliczane są wariancje w obrębie każdego skupienia (średnia odległość między skupieniami). Z tego zestawu liczb (5 w rozwiązaniu 5-klastrowym) wybierana jest minimalna wartość d.
2. Współczynnik transformacji. Według twórców metody wartość Y = P / 2 można przyjąć jako współczynnik transformacji, gdzie P jest wymiarem przestrzeni wektorowej. Jako współczynnik można również przyjąć wartość 1/K, gdzie K jest liczbą skupień.
3. Przekształcone zniekształcenie. Wartość tę oblicza się w następujący sposób:

$$D_t(K) = d^Y(K)$$

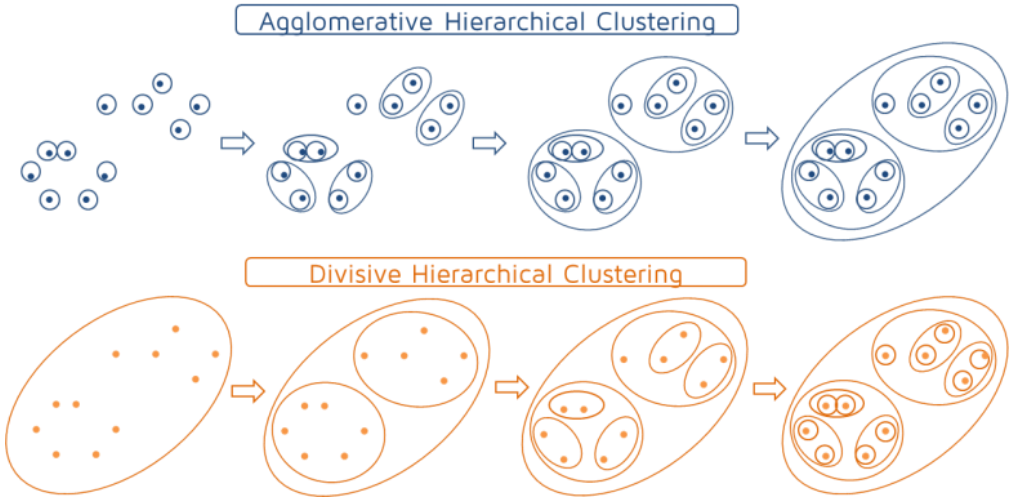


iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the
Erasmus+ Programme
of the European Union

Analiza skupień:
Sugar & James Algorithm

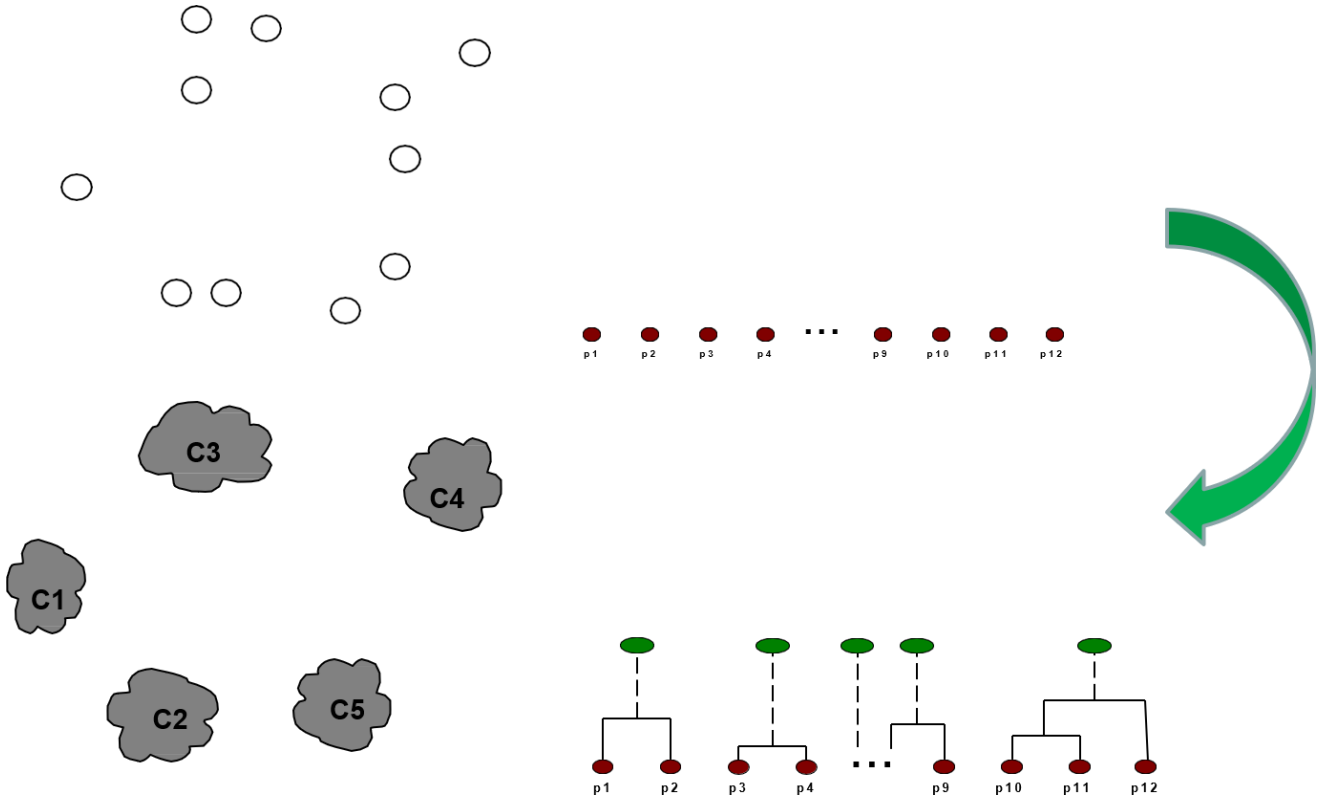
1. Grupowanie aglomeracyjne. Algorytm ten zakłada, że każdy element zbioru jest osobnym klastrem. Proces tworzenia nowych klastrów polega na łączeniu kilku klastrów w jeden nowy klaster. Scalanie odbywa się na podstawie określonej odległości między klastrami. Wykonując takie iteracyjne scalanie, tworzone jest drzewo skupień, które ostatecznie zbiega się w jedno skupienie.
2. Grupowanie dzielące. Ten typ grupowania zakłada, że wszystkie obiekty należą do tego samego klastra. W procesie iteracyjnym klastry są dzielone na kilka różnych klastrów. W związku z tym w tym przypadku powstaje drzewo klastrowe (dendrogram).



iBigWorld:2020-1-PL01-KA203-082197

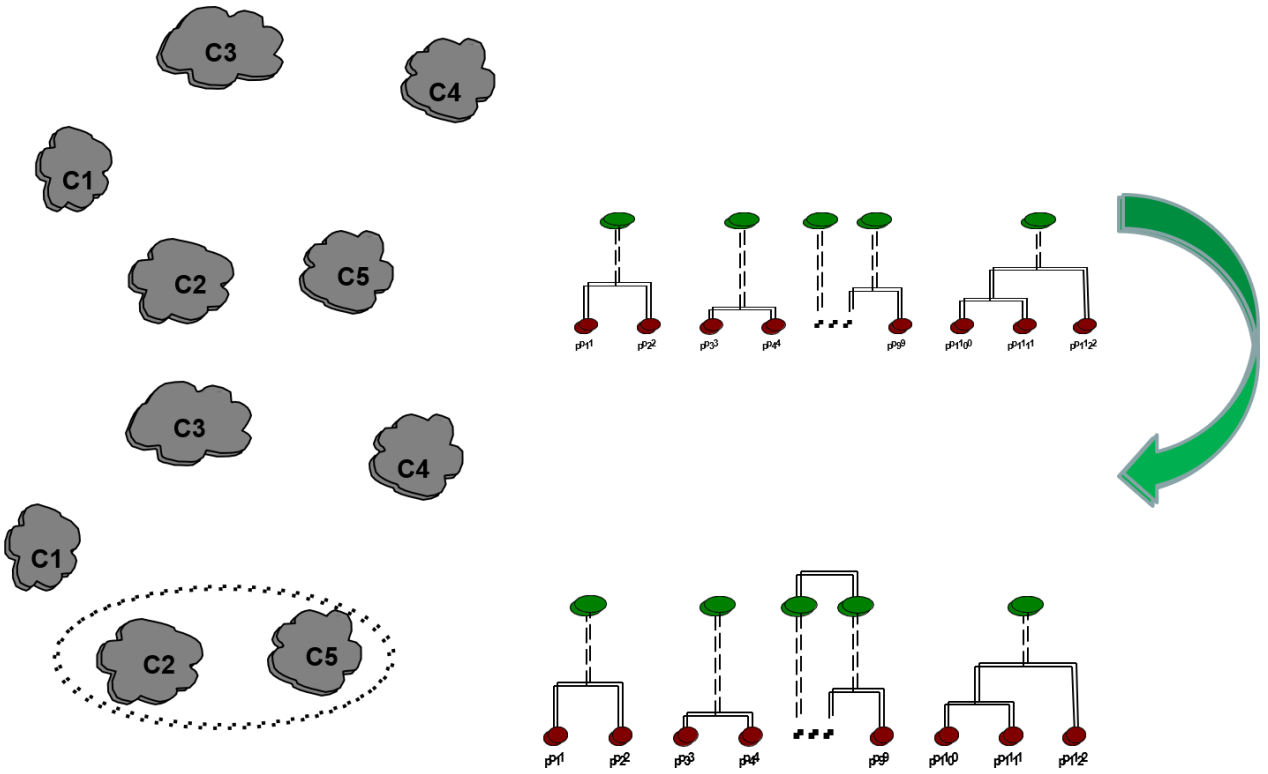
Co-funded by the
Erasmus+ Programme
of the European Union

Analiza skupień:
Średnie grupowanie powiązań



Co-funded by the
Erasmus+ Programme
of the European Union

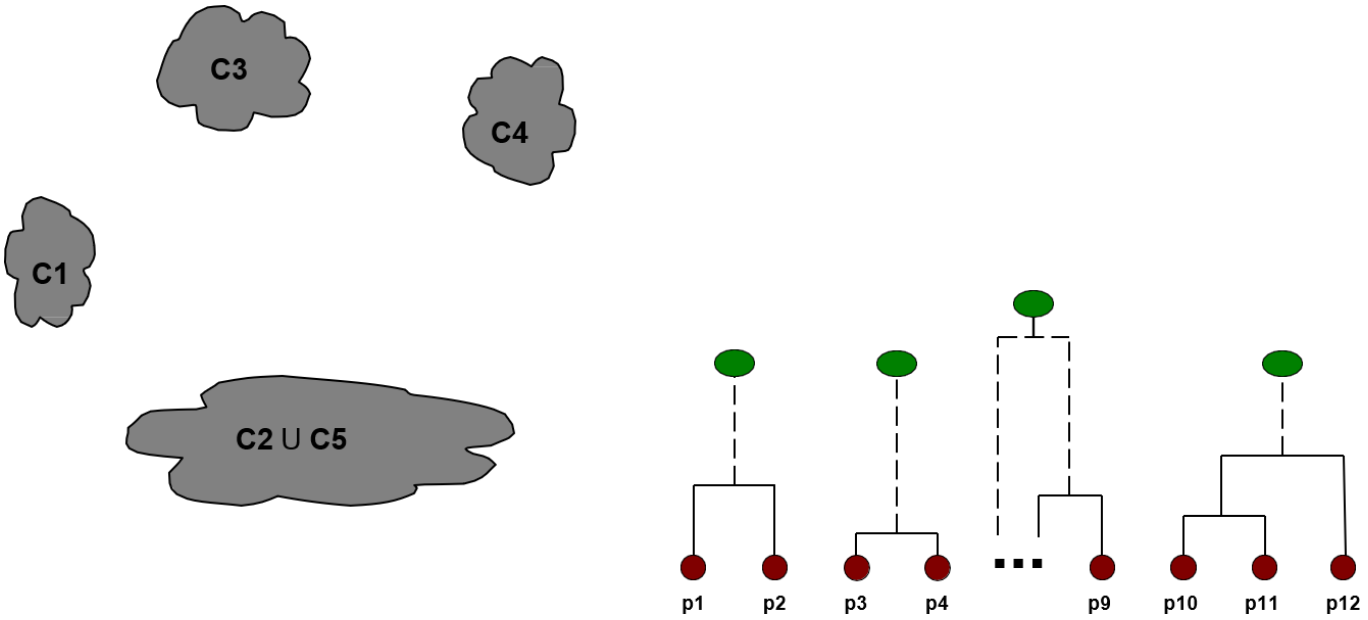
Analiza skupień:
Średnie grupowanie powiązań



iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the
Erasmus+ Programme
of the European Union

Analiza skupień:
Średnie grupowanie powiązań



iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the
Erasmus+ Programme
of the European Union

Analiza skupień: Miary jakości

Można wyróżnić dwa rodzaje mierników jakości:

1. Miary zewnętrzne opierają się na porównaniu automatycznego dzielenia danych z dzieleniem „etalonowym” tych samych danych otrzymanych od ekspertów. Ponadto jako standard można zastosować wartość matematyczną, która jest wybierana na podstawie pewnych podstaw teoretycznych.

Entropia rozwiązania klastra:

$$E(S_r) = -\frac{1}{\log q}\sum_{i=1}^{q}\frac{n_r^i}{n_r}\log\frac{n_r^i}{n_r}, \qquad Entropy = \sum_{r=1}^{k}\frac{n_r}{n}E(S_r).$$

n_r – liczba elementów w danym klastrze,,

q – liczba klas w całej kolekcji,,

n_{ir} – liczba elementów i-tej klasy w klastrze r. Oczywiście, jeśli wszystkie elementy w klastrze należą do tego samego klastra, wtedy log (1) = 0, czyli osiągnięta zostaje minimalna wartość entropii. Zatem chaos (oznaczający brak klastrów) odpowiada maksymalnej wartości entropii, a dobre rozwiązanie klastrów minimum entropii.

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

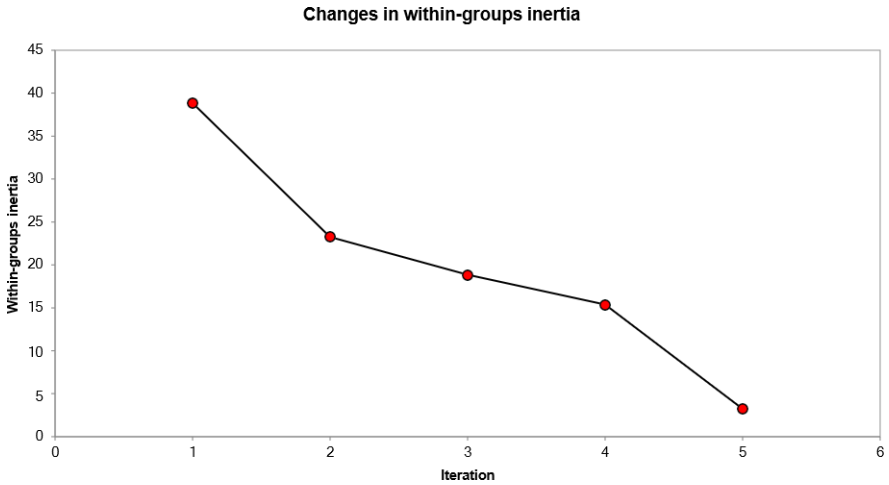
Analiza skupień: Miary jakości

2. Miary wewnętrzne opierają się na ocenie właściwości rozdzielności i zwartości wynikowego podziału danych.

Np. jako miarę używa się sumy kwadratów odchyleń obiektów od środka skupień (bezwładności).

$$\rho_{ij} = \left[\sum_k (x_{ik} - A_k)^2\right]$$

gdzie A_{jk} jest centrum klastra. Podczas iteracji poszukiwana jest minimalna wartość funkcji ρ_{ij}.



iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

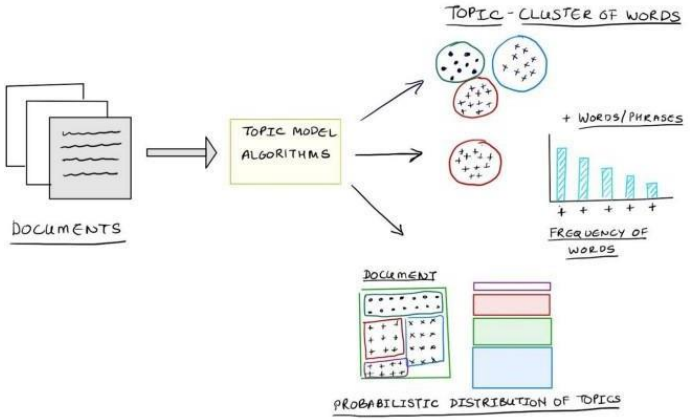
Klasyfikacja i regresja: Modelowanie tematyczne

Modelowanie tematyczne to metoda tworzenia modelu kolekcji dokumentów tekstowych, która określa temat, do którego należą dokumenty.

Model tematyczny zbioru dokumentów tekstowych określa, do jakich tematów należy każdy dokument oraz jakie słowa (terminy) tworzą każdy temat.

Probabilistyczne modele tematyczne są wykorzystywane w następujących obszarach:

1. Wyszukiwanie informacji
2. Identyfikowanie trendów w publikacjach naukowych i wątkach informacyjnych.
3. Klasyfikacja i kategoryzacja dokumentów, obrazów, sygnałów audio i wideo.
4. Określanie tematów społeczności w sieciach społecznościowych.



iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Klasyfikacja i regresja: Modelowanie tematyczne

Zastosowanie modeli tematycznych pozwala uzyskać odpowiedzi na szereg nietrywialnych pytań.

1. Jak rozpoznać sens lub przedmiot dokumentów po ich treści?
2. Jak klasyfikować dokumenty na podstawie tych ukrytych wzorców tematycznych?
3. Jak identyfikować trendy w rozwoju obszarów naukowych?
4. Jak zidentyfikować role ludzi w sieciach społecznościowych?
5. Jak indeksować i automatycznie opisywać dokumenty?
6. Jak przeprowadzić rozpoznawanie wzorców?

Modele tematyczne oparte są na teorii prawdopodobieństwa, w szczególności na zasadzie Bayesa.

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Klasyfikacja i regresja: Teoria prawdopodobieństwa

Różnice w podejściach do teorii prawdopodobieństwa:

Zmienna losowa to wielkość, która w wyniku eksperymentu przyjmuje jedną ze zbioru wartości, a wystąpienia tej lub innej wartości tej wielkości przed jej pomiarem nie można dokładnie przewidzieć.

W podejściu częstotliwościowym (podejście klasyczne) zakłada się, że losowość jest obiektywną niepewnością. Prawdopodobieństwo jest obliczane na podstawie serii eksperymentów i reprezentuje miarę losowości jako empiryczną pewność. Historycznie podejście oparte na częstotliwości wywodziło się z praktycznego problemu: analizy hazardu – dziedziny, w której pojęcie serii testów ma proste i jasne znaczenie.

Podejście bayesowskie zakłada, że niewiedzę charakteryzuje losowość. Na przykład losowość rzutu kostką wiąże się z nieznajomością dynamicznej charakterystyki kości, oporu powietrza i tak dalej. Nie da się rozwiązać wielu problemów metodą częstotliwościową (dokładniej, prawdopodobieństwo pożądanego zdarzenia jest ściśle równe zeru). Jednocześnie interpretacja prawdopodobieństwa jako miary ignorancji pozwala na uzyskanie sensownej odpowiedzi innej niż zero**.**

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Klasyfikacja i regresja: Teoria prawdopodobieństwa

Prawdopodobieństwo wystąpienia zdarzenia A jest stosunkiem liczby skutków sprzyjających temu zdarzeniu do łącznej liczby wszystkich równie możliwych niespójnych skutków elementarnych.

Na przykład. Prawdopodobieństwo, że na kostkę padnie liczba parzysta, jest równe następujący stosunek P = 3/6 = 1/2.

Warunkowe prawdopodobieństwo wystąpienia zdarzenia A, pod warunkiem zajścia zdarzenia B, jest liczbą P(A | B) = P(B, A) / P(B), a P(B, A) jest iloczynem prawdopodobieństw, P(B) to prawdopodobieństwo wystąpienia zdarzenia B.

Na przykład. Załóżmy, że urna zawiera 3 białe i 3 czarne kule. Jedna piłka na raz jest wyjmowana z urny dwukrotnie bez odkładania jej z powrotem. Znajdź prawdopodobieństwo wystąpienia białej kuli w drugim teście (zdarzenie B), jeśli czarna kula została wyciągnięta podczas pierwszego testu (zdarzenie A).

Prawdopodobieństwo zdarzenia A = 3/6 = 1/2.

Iloczyn prawdopodobieństw P (B, A) = (3/6) * (3/5) = 9/30. Wynik końcowy: (9/30) / (1/2) = 3/5.

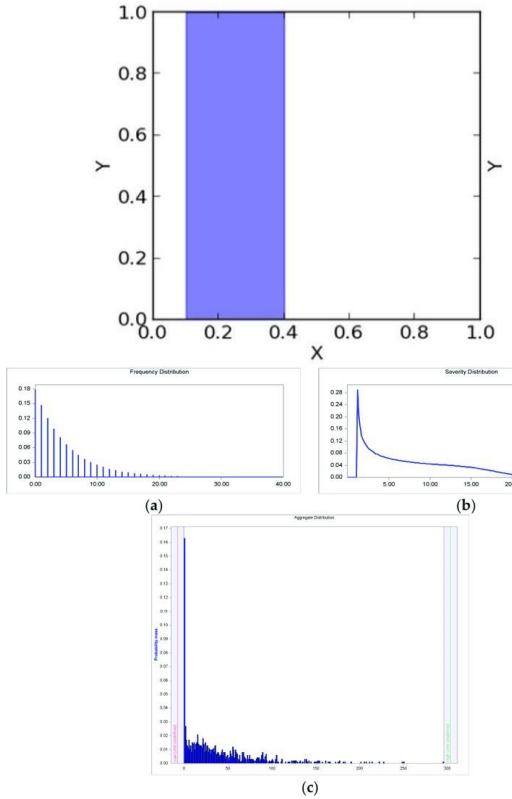
iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Klasyfikacja i regresja: Teoria prawdopodobieństwa

Interpretacja geometryczna rachunku prawdopodobieństwa.

Rozważmy następujący eksperyment: nazwana jest dowolna liczba z przedziału [0, 1] i należy zaobserwować, czy ta liczba będzie zawierała się między np. 0,1 a 0,4. Łatwo się domyślić, że prawdopodobieństwo tego zdarzenia będzie równe stosunkowi długości przedziału [0,1, 0,4] do całkowitej długości przedziału [0, 1] (innymi słowy, stosunkowi „liczby" możliwych równo prawdopodobnych wartości do całkowitej „liczby" wartości), czyli (0,4 – 0,1) / (1 – 0) = 0,3, czyli prawdopodobieństwo trafienia przedziału [0,1, 0,4] wynosi 30%.

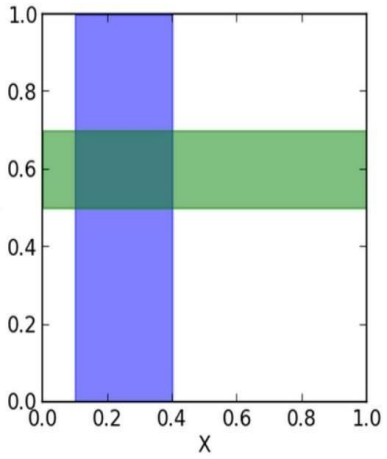


iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Klasyfikacja i regresja: Teoria prawdopodobieństwa

Interpretacja geometryczna rachunku prawdopodobieństwa. Prawdopodobieństwo, że y zawiera się w przedziale [0,5, 0,7] jest równe stosunkowi pola powierzchni zielonej do pola całego prostokąta p (0,5 <= y <= 0,7) = 0,2, czyli w skrócie p (Y) = 0,2. Załóżmy, że konieczne jest znalezienie, jakie jest prawdopodobieństwo, że y jest w przedziale [0,5, 0,7], jeśli x jest już w przedziale [0,1, 0,4]. Prawdopodobieństwo to można zapisać jako p(Y | X). Oczywiście prawdopodobieństwo to jest równe stosunkowi pola ciemnej powierzchni (przecięcie zielonego i niebieskiego powierzchnie - p (X, Y)) do pola powierzchni niebieskiej.



iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Klasyfikacja i regresja: Teoria prawdopodobieństwa

Formuła Bayesa.

Prawdopodobieństwo bayesowskie to interpretacja pojęcia prawdopodobieństwa używanego w teorii bayesowskiej. Prawdopodobieństwo definiuje się jako stopień pewności, że osąd jest prawdziwy.

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

P(A) — prawdopodobieństwo a priori hipotezy A (z góry znane prawdopodobieństwo);
P (A|B) — prawdopodobieństwo hipotezy A po wystąpieniu zdarzenia B (prawdopodobieństwo a posteriori);
P(B|A) — prawdopodobieństwo wystąpienia zdarzenia B, jeśli hipoteza A jest prawdziwa; P(B) — całkowite prawdopodobieństwo wystąpienia zdarzenia B;
P(A|B)) — prawdopodobieństwo wystąpienia zdarzenia A, jeśli hipoteza B jest prawdziwa.
Formuła bayesowska pozwala na „transpozycję przyczyny i skutku": poprzez znany fakt zajścia zdarzenia, powala obliczyć prawdopodobieństwo, które jest ono następstwem danej przyczyny. W ten sposób formuła Bayesa może być wykorzystana do opracowania algorytmów klasyfikacji.

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Klasyfikacja i regresja: Teoria prawdopodobieństwa

Przykład: losowy pacjent uzyskał pozytywny wynik testu na COVID-19. Niech dokładność testu wyniesie 99,8% (tzn. daje wynik pozytywny dla 0,2% zdrowych osób). Jakie jest prawdopodobieństwo, że ten pacjent ma COVID-19?

Prawdopodobieństwo pierwszeństwa: P(chorych) - odsetek chorych w kraju (niech będzie 0,3%).

$$P(sick|test+) = \frac{P(test+|sick) \cdot P(sick)}{P(test+|sick) \cdot P(sick) + P(test+|healthy) \cdot P(healthy)} = \frac{1 \cdot 0{,}003}{1 \cdot 0{,}003 + 1 \cdot 0{,}002} = 60\%$$

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Klasyfikacja i regresja : Stwierdzenia a priori i aposteriori

Załóżmy, że trzeba znaleźć wartość jakiejś nieznanej wielkości.
Mamy pewną wiedzę zdobytą przed (a priori) obserwacją/eksperymentem. Może to być doświadczenie przeszłych obserwacji, niektóre hipotezy modelowe, oczekiwania.
W procesie obserwacji wiedza ta jest stopniowo udoskonalana. Po (a posteriori) obserwacji/eksperymencie powstaje nowa wiedza o zjawisku.
Załóżmy, że konieczne jest oszacowanie nieznanej wartości wielkości P(A|B) poprzez obserwację niektórych jej cech pośrednich (hipotez).

Formuła bayesowska (1763) wyznacza reguły, według których następuje transformacja wiedzy w procesie obserwacji.

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Klasyfikacja i regresja: probabilistyczna definicja problemu klasyfikacji

Niech będzie dany zbiór obiektów X i skończony zbiór klas Y. Wymagane jest zbudowanie algorytmu zdolnego do klasyfikowania losowego obiektu X w obrębie danego zbioru Y. Prawdopodobieństwo a posteriori obiektu X należącego do klasy Y zgodnie z formułą bayesowska wynosi:

$$P(A|B) = \frac{p(A,B)}{P(A)} = \frac{p(A) \cdot P(B|A)}{p(A)},$$

P(A|B) – prawdopodobieństwo a posteriori;

P(A, B) – prawdopodobieństwo a priori.

Problem klasyfikacji polega na obliczeniu (ocenie) informacji a posteriori na podstawie informacji a priori. Taką ocenę można przeprowadzić za pomocą formuły bayesowskiej. Pojawia się jednak problem z oceną wcześniejszej wartości p(A, B).

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Klasyfikacja i regresja: Problem z rekonstrukcją wcześniejszej dystrybucji

p(x,y)

1. Funkcję p(x,y) można oszacować trzema metodami.
2. Nieparametryczna rekonstrukcja gęstości oparta jest na lokalnym przybliżeniu gęstości p(x) w otoczeniu klasyfikowanego obiektu x ∈ X. Przykład: algorytm Parzena-Rosenblatta (metoda okna Parzena).
3. Parametryczna rekonstrukcja gęstości oparta jest na założeniu, że gęstość rozkładu jest znana i odpowiada parametrowi, p(x, y) = ϕ(x; θ), gdzie ϕ jest funkcją stałą. Przykład: Normalna Analiza Dyskryminacyjna, Liniowa Analiza Dyskryminacyjna - oparta na metodzie dekompozycji SVD.
4. Rekonstrukcja kombinacji gęstości. Jeżeli funkcji gęstości p(x, y) nie można zamodelować rozkładem parametrycznym, można ją opisać kombinacją kilku rozkładów:

$$p(x) = \sum_{j=1}^{k} w_j \varphi(x; \theta_j), \sum_{j=1}^{k} w_j = 1.$$

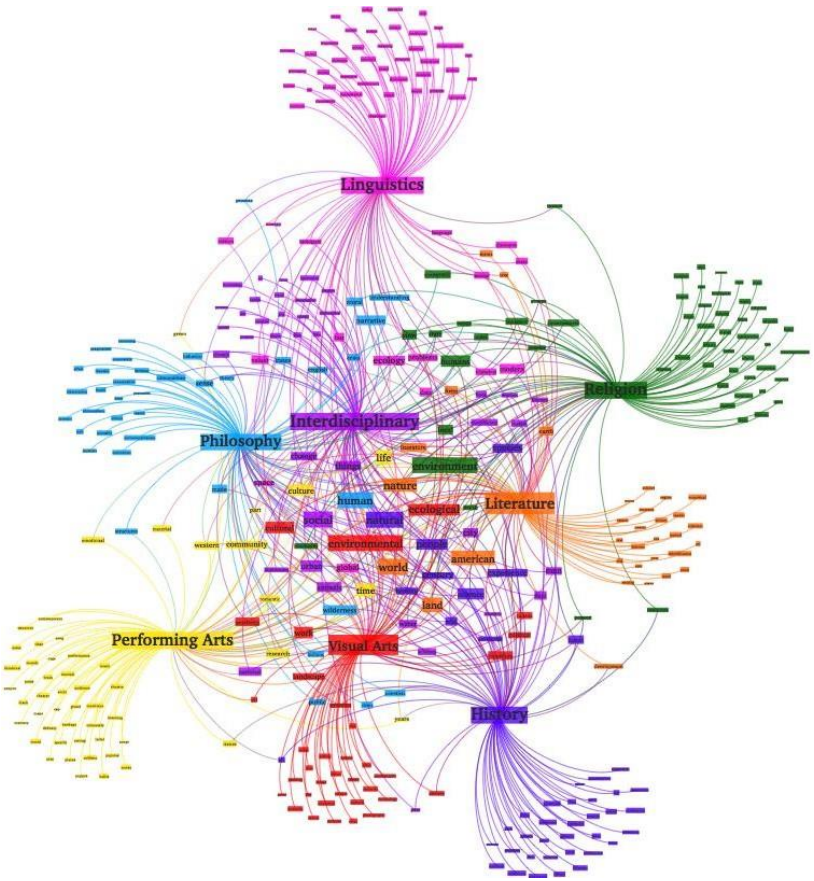
Trzecia metoda to podstawa modelowania tematycznego.

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Klasyfikacja i regresja: modelowanie tematyczne

Temat.
W literaturze dotyczącej modeli tematycznych pojęcie tematu jest definiowane na różne sposoby, w zależności od szkoły tradycji intelektualnej: „ukryte wzorce”, „zwarte opisy sensu dokumentów”, „probabilistyczne (rozmyte) skupiska powiązanych semantycznie terminów ", "powiązanie między terminami a innymi obiektami (dokumentami, autorami, organizacjami, konferencjami itp.), które pozwala na znalezienie ukrytych powiązań skojarzeniowych między nimi "..

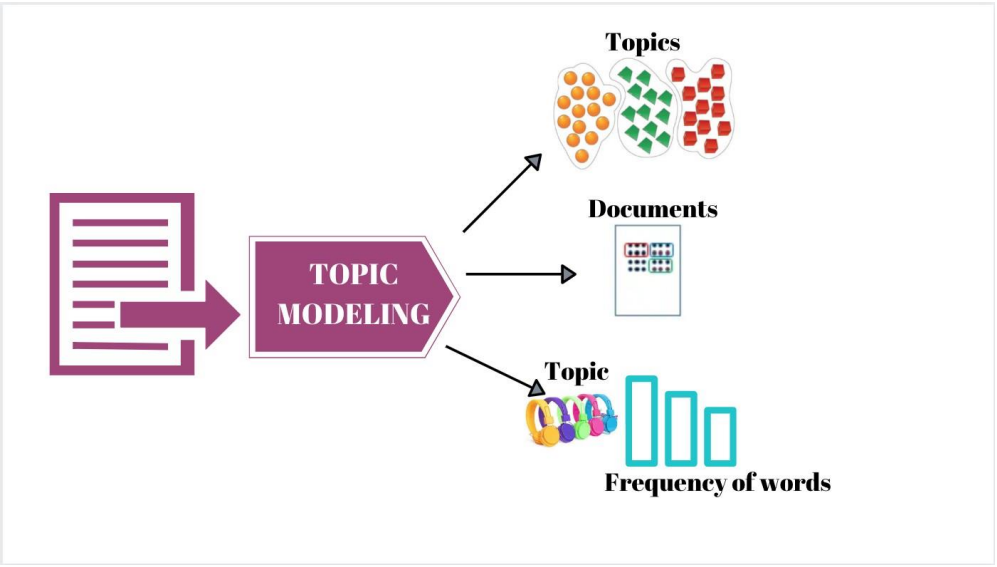


iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Klasyfikacja i regresja: modelowanie tematyczne

Model tematyczny —
to model kolekcji dokumentów tekstowych, który definiuje, do jakich tematów należy każdy dokument w kolekcji. Algorytm budowania modelu tematycznego przyjmuje jako dane wejściowe zbiór dokumentów tekstowych. Na wyjściu dla każdego dokumentu podawany jest wektor liczbowy, złożony z oszacowań stopnia przynależności danego dokumentu do każdego z tematów.



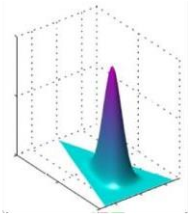
iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Klasyfikacja i regresja: modelowanie tematyczne

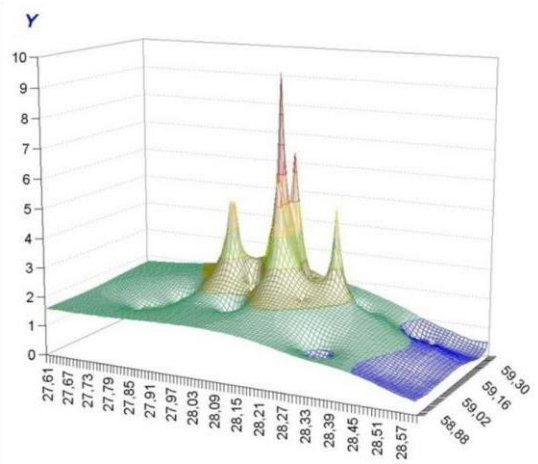
Prawdopodobieństwo, że słowo w wyodrębnione z dokumentu d należy do tematu T, można opisać funkcją złożoną (iloczyn funkcji Dirichleta).

$$P(w \mid t) =$$



Dirichlet function

$$P(w \mid d) =$$



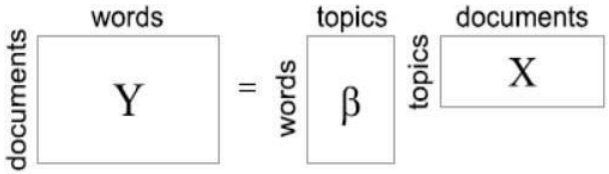
iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Klasyfikacja i regresja: modelowanie tematyczne

Modelowanie tematyczne w zakresie analizy macierzy polega na aproksymacji bardzo dużej macierzy Dokumenty - Słowa to dwie małe macierze:

1. Słowa – Macierz tematów.
2. Dokumenty – Matryca tematów.



$$F[documents \times words] = \Phi[topics \times words] \cdot \Theta[documents \times topics]$$

Problem z przybliżeniem : $F = \Theta \cdot \Phi = (\Theta \cdot R) \cdot (R^{-1}\Phi) = \Theta' \cdot \Phi'$

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Klasyfikacja i regresja: Latent Dirichlet Allocation (LDA)

Podstawowym założeniem modelu tematycznego Latent Dirichlet Allocation jest to, że każdy dokument może z pewnym prawdopodobieństwem należeć do wielu tematów. Temat to zbiór słów, w którym każde słowo ma pewne prawdopodobieństwo przynależności do danego tematu.

Formalnie temat definiuje się jako dyskretny (wielomianowy) rozkład prawdopodobieństwa w

przestrzeń słów danego słownika.

Modelowanie tematyczne jest rozwiązaniem problemu, odwrotnym do klasyfikacji. Każdy dokument w zestawie tekstowym jest traktowany jako obserwowalna losowa, niezależna próbka słów (worek słów) generowana przez jakiś ukryty (ukryty) zestaw tematów. Na podstawie tych danych należy zrekonstruować rozkłady prawdopodobieństwa wszystkich tematów w zestawie i określić, który podzbiór tematów wygenerował każdy dokument.

Modelowanie tematyczne opiera się na zastosowaniu formuły Bayesa, w której rozkład słów i tematów jest wyrażony jako kombinacja gęstości rozkładów słów i dokumentów.

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Klasyfikacja i regeresja: Latent Dirichlet Allocation (LDA)

$$P(z_i = j \mid w_i = m, z_{-i}, w_{-i}) \approx \frac{C^{WT}_{m,j} + \beta}{\sum C^{WT}_{m',j} + V\beta} \cdot \frac{C^{DT}_{d,j} + \alpha}{C^{DT}_{d,j'} + \alpha T}$$

$C^{WT}_{m,j}$ - macierz; komórka: ile razy słowo w jest powiązane z tematem t;

$C^{DT}_{d,j}$ - macierz; komórka: ile razy słowo w w dokumencie d jest powiązane z tematem t;

$\sum_{m'} C^{WT}_{m',j} = n_t$ - wektor; komórka: liczba słów związanych z tematem t;

$C^{DT}_{d,j'} = n_d$ - długość dokumentu d w liczbie słów.

Wyniki symulacji:

1. Macierz rozkładu słów według tematu.

$$\theta_{dj} = \frac{C^{DT}_{d,j} + \alpha}{C^{DT}_{d,j'} + T\alpha}$$

2. Macierz dystrybucji dokumentów według tematu.

$$\phi_{m,j} = \frac{C^{WT}_{m,j} + \beta}{\sum_{m'} C^{WT}_{m',j} + V\beta}$$

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Klasyfikacja i regresja: liczba i jakość tematów

1. Liczba tematów.

Podobnie jak w analizie skupień, pojawia się problem doboru liczby tematów. Ten problem można rozwiązać różnymi metodami, ale nie został w pełni rozwiązany.

A. Zwiększ (zmniejsz) liczbę tematów i obserwuj wynik.

B. Użyj metod, takich jak metoda skoku (podobna do analizy skupień).

2. Jakość tematów.

Problem w tym, że otrzymane tematy mogą być źle zinterpretowane (tzw. śmieciowe tematy). Z reguły, ten problem, można rozwiązać za pomocą analizy eksperckiej. Istnieją techniki, które dobrze sprawdzają się w wyszukiwarkach, na przykład słowa posortowane według prawdopodobieństwa, można mnożyć w każdym temacie przez współczynnik TF-IDF. W związku z tym można obliczyć sumę wartości w każdym temacie, a tym samym podkreślić złe i dobre tematy.

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Klasyfikacja i regresja: miara Kullbacka-Leiblera

Miara Kullbacka-Leiblera jest stosowana, gdy konieczne jest obliczyczenie stopnia podobieństwa między dwoma rozkładami.

$$K = 0.5 \sum_{k}^{W} \Phi_k^1 \log\left(\frac{\Phi_k^1}{\Phi_k^2}\right) + 0.5 \sum_{k}^{W} \Phi_k^2 \log\left(\frac{\Phi_k^2}{\Phi_k^1}\right)$$

Jeśli K = 0 - dwa rozkłady są identyczne.Jeśli K = Max – dwa rozkłady są jak najbardziej różne.Jednak obliczanie w przód miary K-L nie jest wygodne, ponieważ obszar ogona długiego rozkładu ma silny wpływ.

Metryka podobieństwa na podstawie KL:

$$Kn = \left(1 - \frac{K}{Max}\right) * 100$$

Jeśli K_n=100%, to dwa tematy są identyczne.

Jeśli K = 0 tematy są jak najbardziej różne.

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Klasyfikacja i regresja: zastosowanie analizy regresji

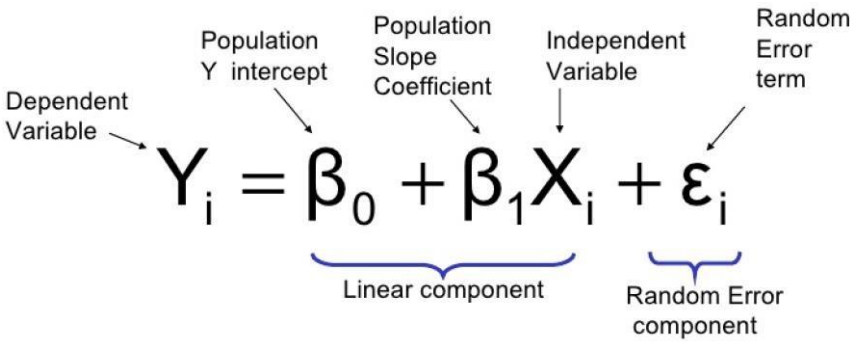
1. Modelowanie liczby zapisów na uniwersytety w celu lepszego zrozumienia czynników przyciągających, które utrzymują dzieci w tej samej instytucji edukacyjnej.
2. Modelowanie przepływów migracyjnych w zależności od czynników takich jak przeciętne wynagrodzenie, dostępność placówek medycznych i szkolnych, położenie geograficzne itp.
3. Symulacja wypadków drogowych w funkcji prędkości, warunków drogowych, pogody, itp.
4. Modelowanie strat spowodowanych wypadkami pożarowymi w funkcji takich zmiennych jak liczba straży pożarnych, czas realizacji zgłoszenia czy cena nieruchomości.

Analiza regresji polega na znalezieniu najważniejszych czynników, które wpływają na zmienną zależną.

Co-funded by the Erasmus+ Programme of the European Union

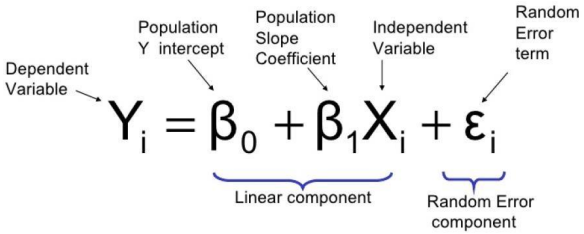
Klasyfikacja i regresja: analiza regresji

Równanie regresji. Jest to wzór matematyczny, który stosuje się do zmiennych niezależnych, aby lepiej przewidywać modelowaną zmienną zależną.



Co-funded by the Erasmus+ Programme of the European Union

Klasyfikacja i regresja: analiza regresji



Zmienna zależna (Y) to zmienna opisująca proces, który należy przewidzieć lub zrozumieć.

Zmienne niezależne (X) to zmienne używane do modelowania lub przewidywania wartości zmiennych zależnych. Znajdują się one po prawej stronie znaku równości w równaniu regresji i są często określane jako zmienne objaśniające. Zmienna zależna jest funkcją zmiennych niezależnych.

Współczynniki regresji (β) to współczynniki obliczane w wyniku analizy regresji. Wartości obliczane są dla każdej zmiennej niezależnej, która reprezentuje siłę więzową i rodzaj relacji zmiennej niezależnej w odniesieniu do zależnej.

Resztki. Istnieje niewyjaśniona liczba zmiennych zależnych przedstawionych w równanie regresji jako błędy losowe ε.

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Klasyfikacja i regresja: analiza regresji

Opracowanie modelu regresji to iteracyjny proces wyszukiwania skutecznych zmiennych niezależnych w celu wyjaśnienia zmiennych zależnych, które należy modelować lub zrozumieć, uruchamiając narzędzie regresji w celu określenia, które wielkości są skutecznymi predyktorami.

Następnie cykl po cyklu usuwając i/lub dodając zmienne, aż do znalezienia najlepiej dopasowanego modelu regresji. Ponieważ proces tworzenia modelu ma często charakter eksploracyjny, nigdy nie może być prostym „dopasowaniem” danych.

Proces tworzenia modelu regresji powinien uwzględniać aspekty teoretyczne, opinie ekspertów terenowych oraz zdrowy rozsądek.

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Klasyfikacja i regresja: analiza regresji

Istnieją regresje liniowe i nieliniowe.
Regresja liniowa:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_{p-1} X_{i,p-1} + \varepsilon_i$$

Regresje nieliniowe dzielą się na dwie klasy: regresje nieliniowe w stosunku do zmiennych objaśniających uwzględnionych w analizie, ale liniowe w ocenianych parametrach oraz regresje nieliniowe w zakresie ocenianych parametrów.

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Klasyfikacja i regresja: analiza regresji liniowej

Opracowanie równania regresji sprowadza się do oceny jego parametrów. Do oceny parametrów regresji, które są liniowe, można zastosować metodę najmniejszych kwadratów. Metoda najmniejszych kwadratów pozwala na uzyskanie takich wartości parametrów, aby suma kwadratów odchyleń cechy efektywnej y wartości rzeczywistych od wartości teoretycznych była minimalna.

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_{p-1} X_{i,p-1} + \varepsilon_i$$

$$\begin{cases} Y^1 = \alpha + \beta_1 X_1^1 + \ldots + \beta_k X_k^1 + u^1, \\ Y^2 = \alpha + \beta_1 X_1^2 + \ldots + \beta_k X_k^2 + u^2, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ Y^n = \alpha + \beta_1 X_1^n + \ldots + \beta_k X_k^n + u^n, \end{cases}$$

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Klasyfikacja i regresja: analiza macierzy i regresji

Model regresji można zdefiniować jako macierz:

$$0 = b_0*1+b_1*(-2),$$
$$1 = b_0*1+b_1*(-1),$$
$$2 = b_0*1+b_1*0,$$
$$3 = b_0*1+b_1*(+1),$$
$$4 = b_0*1+b_1*(+2).$$

$$\begin{bmatrix} 0 \\ 1 \\ 2 \\ 3 \\ 4 \end{bmatrix} = \begin{bmatrix} +1 & -2 \\ +1 & -1 \\ +1 & 0 \\ +1 & +1 \\ +1 & +2 \end{bmatrix} * \begin{bmatrix} b_0 \\ b_1 \end{bmatrix}$$

Y=XB

Macierz odwrotna to macierz A^{-1}, po przemnożeniu której przez pierwotną macierz A dostajemy macierz jednostkową E:

$$AA^{-1} = A^{-1}A = E$$

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Klasyfikacja i regresja: analiza macierzy i regresji

$$\mathbf{YX^{-1}=XX^{-1}B} \quad \Rightarrow \quad \mathbf{YX^{-1}=B}$$

$$\mathbf{XX^{-1}=1}$$

W ten sposób można otrzymać jawną postać układu równań dla składowych wektora B, czyli wymagane rozwiązanie problemu regresji.

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} \cdot \begin{bmatrix} x_{01} & x_{11} & \dots & x_{k1} \\ x_{02} & x_{12} & \dots & x_{k2} \\ \vdots & \vdots & \dots & \vdots \\ x_{0N} & x_{1N} & \dots & x_{kN} \end{bmatrix}^{-1} = \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_k \end{bmatrix}$$

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Klasyfikacja i regresja: Kontekst współczynnika regresji

Ogólnie rzecz biorąc, współczynnik regresji k opisuje, w jaki sposób cecha efektywna (Y) zmieni się w średniej, jeśli cecha czynnikowa (X) wzrośnie o jeden.

Y = 87610 + 2984 X ; X to liczba pracowników; Y to wielkość rocznej produkcji (EUR).

Przykład interpretacji współczynnika regresji:

W równaniu Y = 87610 + 2984 X współczynnik regresji wynosi +2984. Co to znaczy?

W tym przypadku kontekst współczynnika regresji jest taki, że wzrost liczby pracowników o 1 osobę prowadzi do wzrostu rocznej produkcji w średniej o 2984 euro. Właściwości współczynnika regresji:

- Współczynnik regresji może mieć dowolną wartość.
- Współczynnik regresji nie jest symetryczny, tj. zmienia się, gdy pozycje X i Y są zamienione.

Jednostką miary współczynnika regresji jest stosunek jednostki miary Y do jednostki X: miary: ([Y]/[X]).

Współczynnik regresji zmienia się po zmianie jednostek miary X i Y.

Ponieważ efektywna cecha Y jest mierzona w euro, a czynnik czynnikowy X w liczbie pracowników (osób), to współczynnik regresji jest mierzony w euro na osobę (EUR/osobę).

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Klasyfikacja i regresja: współczynnik determinacji R^2

Za główny wskaźnik odzwierciedlający miarę jakości modelu regresji, opisującą relację między zmienną zależną i niezależną modelu, uważany jest z reguły współczynnik determinacji. Współczynnik determinacji opisuje, jaka część zmiennej objaśnianej y jest brana pod uwagę i jest uzależniona od wpływu czynników zawartych w modelu:

$$r^2 = 1 - \frac{\Sigma (y - y')^2}{\Sigma (y - \bar{y}')^2}$$

where y_i - observed variable values,
$\bar{y}$ – mean value over observed data,
$\hat{y}_i$ – model values constructed from the evaluated parameters.

Co-funded by the Erasmus+ Programme of the European Union

Klasyfikacja i regresja: autoregresja

Model autoregresyjny opisuje relację zmiennej Y do samej siebie, a raczej do wartości Y w minionym okresie (dzień, miesiąc, rok itd.).

Standardowa postać modelu autoregresyjnego:
gdzie a_0 jest stałym współczynnikiem opisującym, jaki otrzymamy wynik z modelu w przypadku, gdy wszystkie cechy efektywne są równe zeru;
a_i — współczynniki opisujące stopień zależności końcowej wartości Y od cech efektywnych (w tym przypadku wartości Y w ostatnim okresie regresji);
Y_{i-1} — efektywne cechy, które w tym przypadku są ostatecznymi wartościami Y dla minionego okresu;
ε_i — jest składową losową lub jak to się potocznie nazywa błędem modelu (w rzeczywistości jest to różnica między obliczoną wartością modelu dla znanych okresów a samymi znanymi wartościami, czyli $Y_{obliczone}$ - Y).

Co-funded by the Erasmus+ Programme of the European Union

Wydobywanie wzorów

Eksploracja wzorców to proces identyfikacji reguł mający na celu opisanie określonych wzorców w danych.

Jedną z pierwszych aplikacji do eksploracji danych była analiza koszyka rynkowego, która pozwoliła zidentyfikować pozycje, które zazwyczaj pojawiają się w transakcjach zakupu. Na przykład, jeśli supermarket sprzedaje ryby, to również powinien sprzedawać sos tatarski, ponieważ te dwie pozycje często pojawiały się razem w tych samych zakupach.

Takie asocjacje są zwykle łatwe do odkrycia w małych zestawach danych. Eksploracja danych pozwoliła jednak odkryć mniej oczywiste i nieoczekiwane zależności w dużych zbiorach danych, które są przydatne w różnego rodzaju badaniach.

Ponadto eksploracja wzorców umożliwia wykrywanie wzorców sekwencyjnych, takich jak sekwencje ostrzeżeń i błędów, które mogą być przydatne w zapobieganiu awariom sprzętu.

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Dane szeregów czasowych

W statystyce szereg czasowy oznacza sekwencję danych mierzonych w pewnych (często równych) odstępach czasu. Analiza szeregów czasowych łączy metody badania szeregów czasowych, dążąc zarówno do zrozumienia natury punktów danych, jak i do zbudowania predykcji. Przewidywanie szeregów czasowych polega na budowaniu modelu do przewidywania przyszłych zdarzeń na podstawie znanych zdarzeń z przeszłości, przewidywaniu przyszłych danych przed ich pomiarem. Na przykład przewidywanie ceny otwarcia rynku giełdowego na podstawie jego wcześniejszych działań.

Jakie zadania tu powstają?

- Fizyka słońca: a) okresy ukryte; b) przewidywanie aktywności.
- Elektrokardiogramy (EKG): a) charakter obserwowanych arytmii; b) przewidywanie rozwoju państwa.
- Szeregi ekonomiczne: a) zadanie segmentacji; b) zadanie przewidywania.
- Kinetyka chemiczna: a) analiza dynamiki; b) budowa modelu.

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Dane szeregów czasowych

Trend (tendencja) - stabilny wzorzec obserwowany od dłuższego czasu. Na przykład cechy demograficzne lub wzrost konsumpcji.

Składnik sezonowy to funkcja charakteryzująca wahania sezonowe. Z reguły wahania te są okresowe lub prawie okresowe, np. zatory na drogach, szczyty sprzedaży.

Składnik cykliczny to funkcja opisująca długie okresy (ponad rok) malejących lub rosnących. przykładem takich wahań są fale Kondratiewa, pułapki demograficzne.

Składnik losowy to szum, który odzwierciedla losowe działania wielu czynników.

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the
Erasmus+ Programme
of the European Union

Dane szeregów czasowych: przegląd modeli

1. Modele przewidywania regresji:
• Regresja liniowa.
• Regresja wielokrotna.
• Regresja nieliniowa.

2. Modele predykcji autoregresyjnej:
• ARIMAX (rozszerzona zintegrowana średnia ruchoma autoregresji).
• GARCH (uogólniona autoregresyjna warunkowa heteroskedastyczność).
• ARDLM (model rozproszonych opóźnień autoregresji)

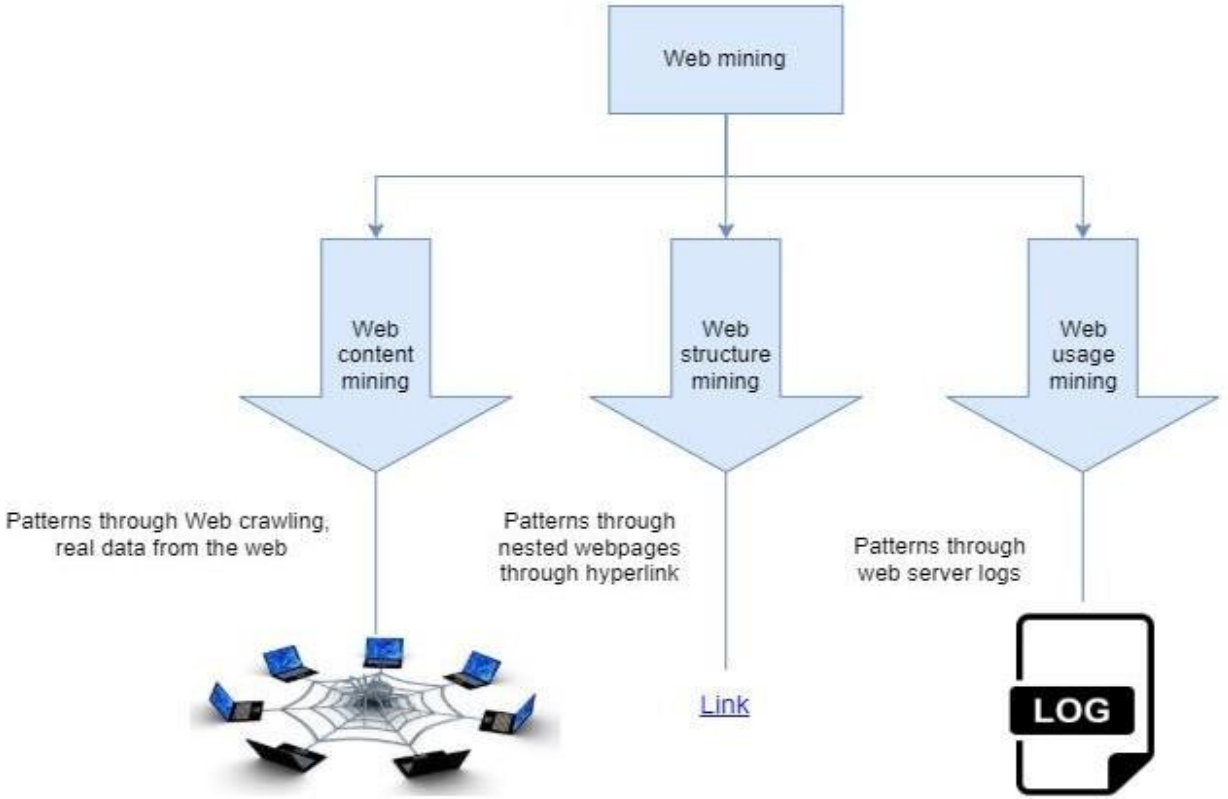
3. Modele wygładzania wykładniczego (ES):
• Ważona średnia krocząca.
• Wygładzanie wykładnicze (model Brown).
• Model Holta.
• Model Holta-Wintersa.

4. Model próbkowania maksymalnego podobieństwa (MMSP).
5. Model sztucznych sieci neuronowych (SSN).
6. Model łańcuchów Markowa.
7. Drzewo Klasyfikacji i Regresji (CART).
8. Model algorytmu genetycznego (GA).
9. Wspieraj model maszyny wektorowej (SVM).
10. Model funkcji transferu (TF).
11. Model logiki rozmytej (FL).
12. Model analizy widma osobliwego (SSA).
13. Lokalny model aproksymacyjny (LA)
14. Fraktalny model szeregów czasowych.
15. Model przekształcenia falkowego.
16. Model transformacji Fouriera.

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

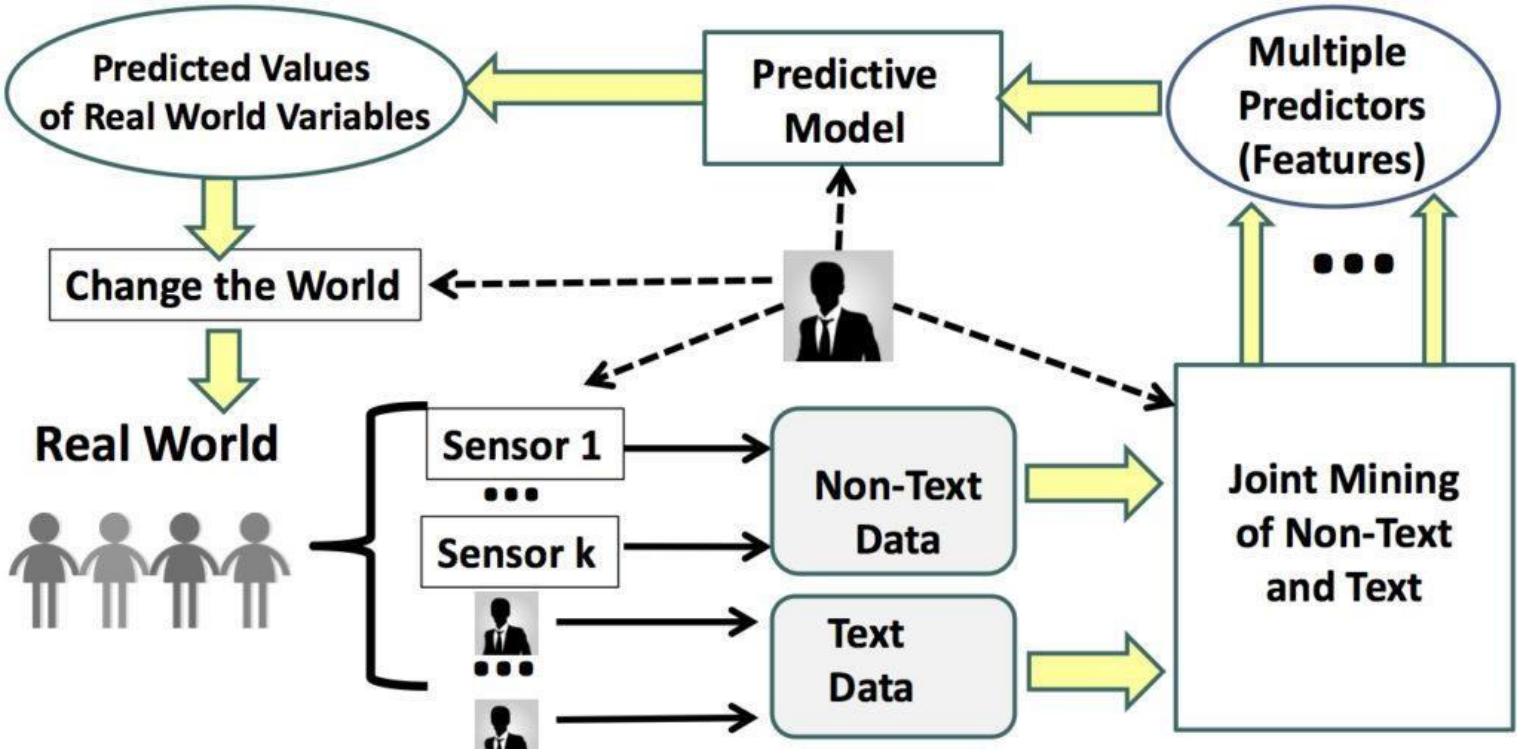
Eksploracja danych WEB



iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Information Retrieval



iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Literatura

1. B. B. Gupta, D. Perakovic, A. A. Abd El-Latif and D. Gupta, “Data Mining Approaches for Big Data and Sentiment Analysis in Social Media”, IGI Global, December 2021, 313 p.
2. J. Leskovec, A. Rajaraman, J. D. Ullman, “Mining of Massive Datasets”, Cambridge University Press; 3rd edition, February 13, 2020, 565 p.
3. A. J. Gutman, J. Goldmeier, “Becoming a Data Head: How to Think, Speak, and Understand Data Science, Statistics, and Machine Learning”, Wiley; 1st edition, April 13, 2021, 240 p.
4. K. Jamsa, “Introduction to Data Mining and Analytics”, Jones & Bartlett Learning, February 17, 2020, 668 p.
5. J. Dean, “Big Data, Data Mining, and Machine Learning: Value Creation for Business Leaders and Practitioners”, Wiley; 1st edition, May 7, 2014, 259 p.

Co-funded by the Erasmus+ Programme of the European Union