Politechnika Śląska

Wydział Automatyki, Elektroniki I Informatyki

Praca Doktorska

Grupowanie danych z wykorzystaniem mieszanin wielowymiarowych rozkładów

Streszczenie

mgr Mateusz Kania

promotor: prof. zw dr hab inż Andrzej Polański

Poniższa praca dyplomowa została wsparta finansowo z funduszy Unii Europejskiej projektu AIDA (Applied Integrative Data Analysis) POWR.03.02.00-00-I029.

Strona celowo pozostawiona pusta

Contents

1 Wprowadzenie

Nienadzorowane uczenie maszynowe jest szeroko stosowaną techniką w analizie danych i uczeniu maszynowym. Celem jest pogrupowanie podobnych obiektów lub obserwacji w grupy bez wcześniejszej znajomości prawdziwych etykiet klas. Może to być trudne zadanie, zwłaszcza gdy mamy do czynienia z wielowymiarowymi i złożonymi zbiorami danych. Wiele różnych algorytmów i podejść może być wykorzystanych do wykonania bezobsługowego grupowania. Jednym ze sposobów kategoryzacji algorytmów grupowania nienadzorowanego jest model i założenia, na którym się opierają.

Jedną z takich grup są algorytmy oparte na modelach, które zakładają, że dane są generowane z mieszaniny rozkładów prawdopodobieństwa. Algorytmy te mają na celu oszacowanie parametrów tych rozkładów, aby wyłonić grupy. Przykładem takiego algorytmu jest algorytm Expectation-Maximization (EM).

1.1 Dostępność codu na GitHub

Kod jest dostępny na githubie: https://github.com/callimae/multivarEM

Aby zainstalować pakiet, należy zainstalować R. Jest to darmowe, open-source'owe środowisko programistyczne, dostępne na stronie: https://www.r-project.org/

Polecenia do wykonania po otwarciu R:

```
install.packages("devtools")
install_github("callimae/multivarEM")
```

2 Algorytmy bazujące na modelach

EM jest potężną techniką obliczeniową stosowaną do estymacji parametrów modelu statystycznego w obecności brakujących lub niekompletnych danych. Algorytm składa się z dwóch naprzemiennych etapów: etapu E i etapu M.

W kroku E algorytm oblicza wartość oczekiwaną ukrytychdanych, biorąc pod uwagę dane obserwowane i aktualną estymację parametrów. Krok ten polega na obliczeniu rozkładu potomnego brakujących danych z wykorzystaniem aktualnej estymacji parametrów.

W kroku M algorytm aktualizuje oszacowania parametrów w oparciu o wartości oczekiwane brakujących danych obliczone w kroku E. Krok ten polega na znalezieniu maksymalnego prawdopodobieństwa estymacji parametrów przy oczekiwanych wartościach brakujących danych.

Algorytm EM jest stosowany w różnych dziedzinach, wizji komputerowej, przetwarzaniu języka naturalnego i bioinformatyce. Jest on skuteczny, gdy dane są niekompletne lub brakuje, a tradycyjne metody maksymalnego prawdopodobieństwa nie mają zastosowania.

Jedną z mocnych stron algorytmu EM jest to, że ma on gwarancję zbieżności do lokalnego maksimum funkcji prawdopodobieństwa i często globalnego maksimum. Algorytm ten jest jednak wrażliwy na wybór początkowych wartości parametrów. Może też być intensywny obliczeniowo, szczególnie w przypadku dużych zbiorów danych. Jest to szczególnie prawdziwe w przypadku danych wielowymiarowych. Aby zminimalizować ciężkie obliczeniowo przypadki, zaimplementowano dwie wersje EM.

2.1 Wielowymiarowa mieszanina Gaussowska EM

Dużą zaletą Gaussian Mixture Model EM uses only diagonal variances of the covariance matrix instead of a full one. Poniżej zaprezentowano końcowe etapty wyprowadzenia.

2.1.1 Warunkowy rozkład zmiennych ukrytych

Przyjmujemy pewne wartości parametrów jako wartości początkowe dla iteracji. Nazywamy to zgadywaniem parametrów. Oznaczamy odgadnięte parametry literą g w indeksie górnym. Korzystając z twierdzenia Bayesa, możemy obliczyć warunkowy rozkład zmiennej ukrytej przy użyciu wylosowanego parametru.

$$p(z_i = k|\mu_k^{\mathrm{g}}, (\Sigma^U)_k^{\mathrm{g}}, \alpha_k^{\mathrm{g}}) = p(k \mid i) = \frac{\alpha_k^{\mathrm{g}} f\left(x_i, \mu_k^{\mathrm{g}}, (\Sigma^U)_k^{\mathrm{g}}\right)}{\sum_{\chi=1}^{K} \alpha_\chi^{\mathrm{g}} f\left(x_i, \mu_\chi^{\mathrm{g}}, (\Sigma^U)_\chi^{\mathrm{g}}\right)} \tag{1}$$

Wiarygodność mnożona jest przez jej prawdopodobieństwo wstępne (proporcję mieszania). Następnie dokonuje się normalizacji poprzez zsumowanie prawdopodobieństwa pomnożonego przez priorytety wszystkich pozostałych składników mieszaniny. Wynikiem jest prawdopodobieństwo potomne z_i wygenerowane przez k komponent mieszaniny.

2.1.2 Warunkowa wartość oczekiwana logarytmicznej funkcji wiarygodności (Q-function)

Używając 1 możemy wyprowadzić równanie warunkowej wartości oczekiwanej funkcji zlogarytmizowanej funkcji wiarygodności. Zgodnie z nomenklaturą często stosowaną w literaturze przedmiotu nazywamy ją funkcją Q.

$$E\left(L^C \mid \text{ data, parameter guess }\right) = Q = \\ = \sum_{i=1}^{N}\sum_{k=1}^{K}\left(\ln \alpha_k\right) p(k \mid i) + \sum_{i=1}^{N}\sum_{k=1}^{K}\left[-\frac{M}{2}\ln(2\pi) - \frac{1}{2}\ln\left|\Sigma_k^U\right| - \frac{1}{2}\left(x_i - \mu_k\right)^T \left(\Sigma^U\right)_k^{-1}\left(x_i - \mu_k\right)\right] p(k \mid i) \tag{2}$$

W równaniu M oznacza liczbę wymiarów.
Maksymalizując funkcję Q w odniesieniu do parametrów α_k,μ_k,Σ_k otrzymujemy:

$$\hat{\alpha}_k = \frac{\sum_{i=1}^{N} p(k \mid i)}{N} \tag{3}$$

$$\hat{\mu}_k = \frac{\sum_{i=1}^{N} x_i p(k \mid i)}{\sum_{i=1}^{N} p(k \mid i)}, \quad k = 1, 2, \ldots, K \tag{4}$$

$$\hat{\Sigma}_k^U = \frac{\sum_{i=1}^{N} [\operatorname{diag}\left(x_i - \hat{\mu}_k\right)]^2 p(k \mid i)}{\sum_{i=1}^{N} p(k \mid i)} \tag{5}$$

gdzie:
$\operatorname{diag}(y) = \begin{bmatrix} y_1 & 0 & \cdots & 0 \\ 0 & y_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & y_M \end{bmatrix}$ - is a diagonal matrix composed with elements of a vector y
$y = [y_1, y_2, ..., y_M]^T$

2.2 Multinomial Mixture EM

2.2.1 Warunkowy rozkład zmiennych ukrytych

Tak jak poprzednio, zgadywane parametry oznaczane są przez "g" w indeksie górnym. Są one potrzebne do zainicjowania iteracji i obliczenia warunkowego rozkładu zmiennej ukrytej za pomocą twierdzenia Bayesa.

$$p(k|d) = \frac{\alpha_k^{\mathbf{g}} \times p_{k_1}^{\mathbf{g},nd_1} \times p_{k_2}^{\mathbf{g},nd_2} \times ... \times p_{k_M}^{\mathbf{g},nd_M}}{\sum_{\kappa=1}^{K} \alpha_\kappa^{\mathbf{g}} p_{\kappa_1}^{\mathbf{g},nd_1} \times p_{\kappa_2}^{\mathbf{g},nd_2} \times ... \times p_{\kappa_M}^{\mathbf{g},nd_M}} = \frac{\alpha_k^{\mathbf{g}} \prod_{m=1}^{M} p_{k_m}^{\mathbf{g},nd_m}}{\sum_{\kappa=1}^{K} \alpha_\kappa^{\mathbf{g}} \prod_{m=1}^{M} p_{\kappa_M}^{\mathbf{g},nd_M}} \tag{6}$$

2.2.2 Warunkowa wartość oczekiwana logarytmicznej funkcji wiarygodności (Q-function)

Funkcja pomocnicza dla wielomianowej mieszaniny wygląda następująco:

$$E\left(L^C \mid \text{ data, parameter guess }\right) = Q = \\ = C + \sum_{d=1}^{D}\left[\sum_{k=1}^{K}\ln\left(\alpha_k\right) p(k \mid d) + \sum_{m=1}^{M}\sum_{k=1}^{K} n_{dm}\ln\left(p_{km}\right) p(k \mid d)\right] \tag{7}$$

Maksymalizując funkcję Q w odniesieniu do parametrów α_k, p_k, otrzymujemy:

$$\hat{\alpha} = \frac{\sum_{d=1}^{D} p(k|d)}{D} \tag{8}$$

gdzie:
$\hat{\alpha}$ - jest wektorem propocji mieszania
D - jest liczbą obserwacji

$$\hat{p}_{km} = \frac{\sum_{d=1}^{D} n_{dm} p(k|d)}{\sum_{n=1}^{M} \sum_{d=1}^{D} n_{dm} p(k|d)} \tag{9}$$

gdzie:
n_{dm} - jest wektorem obserwacji
$\hat{p}_{km}$ - jest wektorem nowych proporcji prawdopodobieństwa

3 Algorytmy bazujące na dystansie

Grupowanie hierarchiczne jest algorytmem grupującym punkty danych w klastry na podstawie podobieństwa. Algorytm tworzy drzewiastą strukturę klastrów zwaną dendrogramem, gdzie każdy węzeł reprezentuje klaster, a liście odpowiadają Istnieją dwa rodzaje grupowania hierarchicznego: aglomeracyjne i rozdzielcze. W aglomeracyjnym grupowaniu hierarchicznym algorytm zaczyna od każdego punktu danych jako jego klastra i łączy najbliższe pary klastrów, aż wszystkie punkty będą należały do jednego klastra. W dzielącym grupowaniu hierarchicznym algorytm zaczyna od wszystkich punktów danych w jednym klastrze i rekurencyjnie dzieli go na mniejsze klastry, aż każdy punkt znajdzie się w swoim klastrze.

K-means to algorytm grupowania oparty na centroidach, który dzieli dane na k klastrów, gdzie k jest liczbą predefiniowanych klastrów. Algorytm rozpoczyna się od losowego przypisania k centroidów do punktów danych. Następnie iteracyjnie przypisuje każdy punkt danych do najbliższej centroidy i aktualizuje centroidy na podstawie nowo utworzonych klastrów, aż algorytm będzie zbieżny.

K-medoidy to odmiana k-średnich, która zamiast centroidów wykorzystuje medoidy, reprezentatywne punkty danych w obrębie każdego klastra. Algorytm rozpoczyna się od losowego wyboru k medoidów z punktów danych. Następnie iteracyjnie przypisuje każdy punkt danych do najbliższego medoidu i aktualizuje medoidy w oparciu o nowo utworzone klastry aż do osiągnięcia zbieżności.

Fuzzy c-means jest miękkim algorytmem grupowania, który przypisuje każdemu punktowi danych wartość przynależności do każdego klastra, wskazując stopień, w jakim punkt należy do każdego klastra. Algorytm losowo przypisuje wartości przynależności do każdego punktu i iteracyjnie aktualizuje wartości przynależności i centroidy klastrów w oparciu o bieżące przypisania aż do zbieżności.

Istotną różnicą pomiędzy grupowaniem hierarchicznym a pozostałymi trzema algorytmami jest to, że grupowanie hierarchiczne tworzy hierarchię grup, podczas gdy pozostałe tworzą stałą liczbę klastrów. Główną różnicą między k-średnimi i k-średnimi jest sposób reprezentowania każdego klastra, odpowiednio za pomocą centroidów lub medoidów. Fuzzy c-means różni się od pozostałych algorytmów przypisywaniem wartości przynależności do każdego punktu, a nie twardym przypisywaniem klastrów.

Algorytmy te między sobą wykorzystują metrykę odległości do pomiaru podobieństwa między punktami danych i przypisania ich do klastrów.

4 Potok zdarzeń

4.1 Gromadzenie danych

Wszystkie zbiory danych można podzielić na dwie części. Pierwsza z nich obejmuje sztucznie stworzone dane, które mogą stanowić bardziej kontrolowane wyzwanie dla działania algorytmu. Drugi podzbiór zawiera rzeczywiste zbiory danych, które zostały wybrane z uwzględnieniem kryterium wielowymiarowości.

Dane	Źródlo
Zliczenia Mutacji Somatycznych	TCGA
Ekspresja Genów	TCGA/cBioportal
Częstotliwość Kodonów	UCI
Aktywność Sportowa	UCI
Archiwum Wolnej Muzyki	GitHub
Artymia	UCI
NASA Kepplers	Kaggle

Table 1: Prawdziwe dane i bazy z których pochodzą

Wielowymiarowe mieszaniny normalne

Wielowymiarowe mieszaniny normalne zostały wygenerowane przy użyciu wektorów losowych (μ), macierzy wariancji (Σ), proporcji mieszania (α) i wymiarów (d). μ pochodziło z rozkładu jednostajnego w zakresie od 0 do 10, natomiast Σ zakładało zerową korelację między zmiennymi. Wariancje pochodziły z rozkładu jednostajnego w zakresie od 0,1 do 1, a α wyciągnięto z rozkładu jednostajnego w zakresie od 0,1 do 1, przy czym wektor proporcji mieszania znormalizowano tak, aby suma była równa 1. Utworzono ponad 15 000 plików z 2 do 10 składowych i 5 do 1600 wymiarów, używając dwóch pakietów PRNG: MASS i MultiRNG. Stworzono ponad 15 000 plików.

Mieszaniny wielomianowe

Mieszaniny wielomianowe były generowane przy użyciu losowych wektorów prawdopodobieństwa (p), liczb obserwacji (n), proporcji mieszania (α) i stałych wymiarów (d). p pochodziło z rozkładu jednostajnego pomiędzy 0 a 1 i znormalizowane do sumy 1. n było generowane z przedziału [3, n], a α było obliczane przy użyciu rozkładu jednostajnego pomiędzy 0.1 a 1, z wartościami znormalizowanymi do sumy 1. Utworzono ponad 8000 plików zawierających od 2 do 10 klastrów i od 5 do 1600 wymiarów.

4.1.1 Prawdziwe dane

Najważniejszą częścią naszej analizy porównawczej algorytmów klasteryzacji bez nadzoru są eksperymenty obliczeniowe dotyczące publicznie dostępnych zbiorów danych. Można wykorzystać obszerne zbiory danych ze świata rzeczywistego, znajdujące się na np. Kaggle, UCI, NCBI czy NCI. Przy wyborze zbiorów danych do klasteryzacji zastosowano następujące kryteria:

- wysoka wymiarowość
- cechy numeryczne (rzeczywiste lub całkowite).
- różnorodność typów danych
- dostęp do prawdziwych etykiet

Zgodnie z wymienionymi kryteriami, dane powinny posiadać wiele wymiarów. Mieszanina nie powinna być jedno- lub dwuwariantowa, a cechy nie mogą być kategoryczne.

Różne typy danych zakładają, że dane pochodzą z różnych dziedzin wiedzy lub są inaczej mierzone. Na przykład, liczba mutacji somatycznych i ekspresja genów pochodzą z tego samego projektu TCGA, ale zostały zmierzone w inny sposób.

Dane rzeczywiste składają się z różnych zbiorów danych z kilku różnych dziedzin, chociaż większość z nich ma jakieś genetyczne i medyczne podstawy. W związku z tym wybrano następujące dane (Tabela 1):

Zliczenia Mutacji Somatycznych

Badanie to skupia się na mutacjach somatycznych DNA u pacjentów z nowotworami, co jest krytycznym tematem w badaniach nowotworów. Mutacje somatyczne mogą wynikać z ekspozycji na mutageny, błędy w replikacji lub inne czynniki środowiskowe. Można je sklasyfikować na dwa główne typy: mutacje kierujące

(driver mutations), które są związane z rozwojem raka, oraz mutacje pasażerskie (passenger mutations), które mają niewielki lub żaden wpływ. Rozróżnienie tych mutacji jest złożone i obejmuje różne czynniki, w tym ich pozycje w DNA, czy wpływ na transkrypcje.

Niniejsze badanie ma na celu grupowanie pacjentów, u których zdiagnozowano różne typy nowotworów, na podstawie zliczeń mutacji somatycznych w genach. Zliczenia mutacji w genach zostały wykorzystane jako obserwacyjne dane wektorowe. Postawiono hipotezę, że informacje te mogą być wykorzystane do rozróżnienia różnych typów nowotworów.

Dane wykorzystane w badaniu zostały pozyskane z The Cancer Genomic Atlas (TCGA) i anotowane przy użyciu callera wariantów somatycznych Mutect2. Przetworzone dane zawierały informacje o numerze próbki, nazwę genu, w którym wystąpiła mutacja oraz częstotliwość mutacji. Badanie miało również na celu określenie, które algorytmy klasteryzacji najlepiej poradziły sobie z zadaniem grupowania. Ogólnie rzecz biorąc, badanie przyczynia się do zrozumienia złożonej natury mutacji somatycznych w nowotworach.

Ekspresja Genów

Ekspresja genów to proces, w którym komórka wykorzystuje informacje zapisane w genomie do tworzenia funkcjonalnych białek, co prowadzi do manifestacji obserwowalnych cech lub fenotypów z danych genetycznych. Proces ten wymaga udziału enzymów i białek w celu przekształcenia informacji zakodowanej w DNA w cząsteczki RNA, nazywane zbiorczo transkryptomem. Messenger RNA (mRNA) to cząsteczka RNA, która odgrywa krytyczną rolę w tłumaczeniu informacji genetycznej na białka. Podczas translacji sekwencja mRNA jest dekodowana na aminokwasy, które są budulcem białek. Wszelkie zmiany w poziomie ekspresji mRNA mogą wpływać na syntezę białek i prowadzić do rozwoju różnych chorób, w tym nowotworów. cBioPortal to platforma internetowa, która zapewnia dostęp do danych molekularnych i klinicznych z różnych badań genomowych nowotworów. Dane dostępne na platformie obejmują metylację DNA, ekspresję mRNA, ekspresję mikroRNA oraz dane dotyczące poziomu białka.

Prezentowana analiza skupiła się na danych dotyczących ekspresji mRNA, aby zbadać, czy ten typ danych może rozróżnić różne rodzaje nowotworów na podstawie ich wzorców ekspresji. Dane o ekspresji genów na cBioPortal są wielowymiarowe, zawierają informacje o licznych genach w wielu typach nowotworów. Użyto techniki transpozycji macierzy do przeformatowania danych, co pozwoliło nam przedstawić numery próbek jako obserwacje, a geny jako zmienne. Pozwoliło nam to na zbadanie zależności pomiędzy różnymi typami nowotworów w oparciu o ich wzorce ekspresji genów, dostarczając cennego wglądu w biologię nowotworów i potencjalne cele terapeutyczne. Pełne dane składały się z ponad 9 000 obserwacji i około 35 000 cech, podobnie jak w przypadku Zliczeń mutacji somatycznych.

Częstotliwość Kodonów

Kod genetyczny tłumaczący sekwencje DNA na aminokwasy był długo uważany za uniwersalny dla wszystkich gatunków. Jednak ostatnie odkrycia pokazały, że istnieją wyjątki od tej reguły, szczególnie w przypadku niestandardowych kodonów używanych przez genomy mitochondrialne. Zrozumienie częstotliwości użycia kodonów w różnych organizmach może dostarczyć wiedzy na temat ich składu genetycznego i tożsamości taksonomicznej. Zbiór danych o częstotliwości użycia kodonów w różnych organizmach, pochodzący z bazy Codon Usage Tabulated from GenBank (CUTG), został przeanalizowany przy użyciu algorytmów uczenia nienadzorowanego w celu zbadania, czy struktura gatunków może zostać odtworzona poprzez podział na odpowiednie królestwa w oparciu o częstotliwość użycia kodonów. Zestaw danych został dostarczony przez Bohdana Khomtchouka z Uniwersytetu w Chicago i zawierał częstotliwości różnych kodonów używanych przez gatunki w kilku królestwach. Wyniki sugerowały, że częstotliwości wykorzystania kodonów mogą być pomocnym narzędziem do rozróżniania gatunków i identyfikacji ich przynależności taksonomicznej. Stwierdzono jednak również, że skuteczność tego podejścia różni się w zależności od konkretnego zastosowanego algorytmu.

Badanie podkreśla znaczenie rozważenia różnych metod podczas analizy danych dotyczących użycia kodonów i podkreśla wartość technik uczenia maszynowego dla zrozumienia zjawisk biologicznych.

Aktywność sportowa

Popularność smartwatchów znacząco wzrosła w ostatnich latach, a wiele z nich oferuje szereg funkcji takich jak śledzenie tętna, ilość spalonych kalorii oraz różne statystyki związane z aktywnością sportową. Bardziej zaawansowane smartwatche mogą automatycznie wykrywać działania użytkownika i rejestrować odpowiednie metryki. W badaniu dotyczącym aktywności sportowych, przeanalizowano dane udostępnione przez profesora z Yale, Billura Barshana. Dane zawierały informacje o 19 aktywnościach sportowych wykonywanych przez ośmiu uczestników, czterech mężczyzn i cztery kobiety, w wieku od 20 do 30 lat. Nie otrzymali oni wcześniej żadnych instrukcji dotyczących wykonywania ćwiczeń. Dane zostały zebrane za pomocą czujników umieszczonych na różnych częściach ciała, z częstotliwością próbkowania 25 Hz, co dało 45 atrybutów i 1140000 obserwacji.

Archiwum Wolnej Muzyki

Free Music Archive to kompleksowy projekt analizy muzyki, który oferuje ogromną bibliotekę 106 577 utworów obejmujących 16 341 artystów na 14 854 albumach. Dane, które zostały udostępnione publicznie pod koniec 2016 roku, obejmują: wstępnie obliczone cechy audio. Metadane na poziomie użytkownika i utworu. Pełnowymiarowe, wysokiej jakości audio dla wybranych gatunków. Cechy audio zostały obliczone przy użyciu librosa, bogatej biblioteki Pythona do analizy audio i muzyki, która umożliwia niskopoziomową ekstrakcję cech, takich jak chromagramy, spektrogramy Mel, Mel Frequency Cepstral Coefficient (MFCC) oraz inne cechy spektralne i rytmiczne. Wszystkie utwory w archiwum są zorganizowane w hierarchiczną taksonomię 161 gatunków, w tym rock, jazz, pop i muzyka klasyczna. Dane zostały znacznie zredukowane do dwunastu najczęściej występujących gatunków na potrzeby analizy, co dało ostateczny zbiór danych składający się z 49 598 obserwacji z 518 cechami.

Arytmia

Arytmia jest rodzajem choroby serca, gdzie bicie serca jest nieregularne, zbyt szybkie lub zbyt wolne. Migotanie przedsionków, specyficzny rodzaj arytmii, jest znany z tego, że zwiększa ryzyko udaru nawet pięciokrotnie. Migotanie przedsionków jest odpowiedzialne za prawie 20% do 30% udarów, a te udary są często bardziej poważne i śmiertelne niż te spowodowane przez inne czynniki. Co szokujące, udary spowodowane migotaniem przedsionków prowadzą do śmierci znacznie częściej niż udary spowodowane innymi przyczynami. Według badań przeprowadzonych w 2016 roku, prawie 7,6 mln osób w Unii Europejskiej miało migotanie przedsionków. Przewiduje się, że do 2060 roku liczba ta wzrośnie o 89% do 14,4 mln, przy czym obecna częstość występowania wzrośnie o 22% z 7,8% do 9,5%. Leczenie arytmii pochłania znaczną część funduszy europejskich wydawanych na ochronę zdrowia, a roczny koszt leczenia wynosi od 0,28% do 2,6% funduszy. Dane analizowane w rozprawie pochodzą z badania, którego celem było rozróżnienie obecności i braku zaburzeń rytmu serca. Obserwacje zostały pogrupowane w szesnaście kategorii, przy czym pierwsza kategoria reprezentowała zwykłe klasy EKG, a pozostałe kategorie różne rodzaje arytmii.

NASA Kepplers

Kosmiczny Teleskop Kepler NASA został wystrzelony 6 marca 2009 roku, aby zidentyfikować inne planety nadające się do zamieszkania poza naszym układem słonecznym. Skupiając się na obszarze około 150 000 gwiazd podobnych do Słońca, Kepler dokonał znaczących odkryć, w tym setek układów gwiezdnych goszczących wiele planet i identyfikując planety, które orbitują w tzw. strefach nadających się do zamieszkania. Panujaca tam temperatura powierzchni może być odpowiednia dla życiodajnej ciekłej wody. Dane numeryczne zmierzone przez teleskop Keppler zostały przeanalizowane za pomocą algorytmów bez nadzoru, aby sprawdzić podobieństwa między planetami.

4.2 Wstępne przetwarzanie danych

Po pobraniu i oczyszczeniu danych, zostały one sparsowane do ustalonego formatu. Dla każdego zestawu danych utworzono macierz o rozmiarze $n \times d$, przy czym obserwacje n umieszczono w wierszach, a zmienne d w kolumnach, aby zapewnić spójność wszystkich analizowanych zestawów danych.

No	Abbreviation	Full name
1	CMN	C-means
2	MANCMN	C-means with Manhattan distance
3	GMM	Gaussian Mixture EM
4	GMMK	Gaussian Mixture Models with k-means
5	HC	Hierarchical Clustering
6	MANHC	Hierarchical Clustering with Manhattan distance
7	KMN	K-means
8	KMN++	K-means++
9	KMD	K-medoids
10	MMM	Multinomial Mixture EM
11	MMMK	Multinomial Mixture EM with k-means

Table 2: Algorithms used in the clustering

Poza zbiorem danych NASA Kepler, pozostałe zbiory danych zawierały zwykle dziesięć lub więcej klas. Metoda permutacji została zastosowana do wygenerowania 10 zestawów danych dla mieszanek 2-, 3-, 4-, 5- i 6-składnikowych w celu stworzenia wielu zestawów danych do analizy. W ten sposób każdy zestaw sześcioskładnikowych mieszanin zawsze zawierał co najmniej jedną lub więcej różnych klas. W rezultacie z jednego rzeczywistego zbioru danych utworzono 50 zbiorów danych z klasami w różnych konfiguracjach.

Metodologia została zastosowana do wszystkich zbiorów danych do analizy.

4.3 Filtracja i skalowanie danych

Brak filtracji

W tym przypadku zbiory danych zawierające wszystkie oryginalne zmienne pozostawiono bez zmian. Zostały one wykorzystane w kroku skalowania oraz na etapie grupowania w analizie.

Dekompozycja wariancji

Obliczano odpowiednią wariancję dla każdej zmiennej, w wyniku czego otrzymywano wektor wariancji o różnych wartościach, zaleznie od użytego zbioru danych. Przyjęto, że otrzymany wektor reprezentuje jednowymiarową mieszaninę składającą się ze zmiennych istotnych i dodatkowego szumu. Do rozróżnienia tych elementów zastosowano technikę dekompozycji mieszaniny przy użyciu pakietu mclust. Używając metryki BIC, wybrano odpowiednią liczbę klastrów. Następnie pozostawiono grupę, która posiadała największą średnią wariancję.

Skalowanie

Skalowanie przeprowadzono osobno na kompletnych i zredukowanych zbiorach danych, aby zachować wariancję. Zmienne skalowano poprzez odjęcie od wektora obserwacji wartości średniej i podzielenie otrzymanych wartości przez odchylenie standardowe. Po skalowaniu dla każdego zbioru danych utworzono dwa dodatkowe pliki.

4.4 Grupowanie danych

Do grupowania danych zastosowano wszystkie wymienione w tabeli algorytmy. Każdy algorytm generował osobny plik, który był zapisywany na dysku w celu późniejszej obróbki.

4.5 Ewaluacja grupowania

W tym badaniu grupy były oceniane na podstawie ich prawidłowego przyporządkowania. W tym celu zastosowano algorytm węgierski do przypisania wyników do etykiet. Następnie zaimplementowano kilka metryk,

aby porównać wydajność algorytmów z różnych perspektyw. Wreszcie, wyniki zostały zwizualizowane przy użyciu różnych typów wykresów.

4.5.1 Przypisanie do grupy - algorithm węgierki (Hungarian)

Algorytm węgierski to algorytm optymalizacji kombinatorycznej, który ma na celu znalezienie optymalnego dopasowania pomiędzy dwoma zbiorami o równej wielkości. Tutaj, został on wykorzystany do znalezienia etykiet zgrupowanych obserwacji.

4.5.2 Metryki oceny jakości klastrów

Do pomiaru jakości grupowania wykorzystano następujące metryki:

- Skorygowany indeks Randa (ARI)
- Indeks Jaccarda i ważony indeks Jaccarda (WJACC)
- Zrównoważony wskaźnik dokładności (BACCU)
- Współczynnik prostego dopasowania (SMC
- Ważony współczynnik prostego dopasowania (WSMC)
- Sprzężony rozkład beta-dwumianowy (BBD)

4.5.3 Wizualizacja

Redukcja wymiarowości

Metody redukcji wymiarowości pozwalają na reprezentację obserwacji w przestrzeni o mniejszej liczbie wymiarów niż w oryginalnych danych. Ułatwia to wizualizację. W tym badaniu zastosowano następujące techniki:

- Analiza głównych składowych (PCA)
- przycięty rozkład wartości pojedynczych (tSVD)
- t-rozproszone stochastyczne osadzanie sąsiadów (tSNE)
- Losowa projekcja (Random projections)

Metryki Do raportowania metryk użyto różnych wykresów.

- ARI - zostało pokazane za pomocą boxplotów.
- BBD - został narysowany z wykorzystaniem funkcji gęstości prawdopodobieństwa beta.
- BACCU - zostało narysowane na płaszczyźnie współrzędnych biegunowych.
- WSMC, SMC i WJACC - te trzy metryki zostały narysowane w postaci wykresów skrzypcowych. Dodatkowo, wykresy zawierają również średnie i pojedyncze wyniki w postaci punktów.
- Wykres korelacji - współczynnik korelacji został obliczony pomiędzy każdą z dwóch metryk, aby zmierzyć siłę i kierunek zależności liniowej w jakości grupowania.

Nr.	Algorytm	Kompletne	Zredukowane	Przeskalowane	Przeskalowane i Zredukowane
1	GMM	1.00	0.79	0.96	0.81
2	KMN	0.81	0.57	0.84	0.58
3	KMN++	0.73	0.51	0.75	0.52
4	GMMK	0.74	0.48	0.71	0.41
5	MANHC	0.67	0.42	0.67	0.42
6	HC	0.63	0.40	0.63	0.40
7	CMN	0.54	0.39	0.56	0.40
8	MMMK	0.59	0.37	0.04	0.00
9	MMM	0.53	0.32	0.04	0.00
10	MANCMN	0.16	0.10	0.15	0.10
11	KMD	0.13	0.05	0.13	0.06

Table 3: Wpisy reprezentują zsumowane sukcesy (poprawnie przypisane klastry), znormalizowane metodą minimum-maksimum w symulowanych wielowymiarowych mieszaninach Gaussowskich

Nr.	Algorytm	Kompletne	Zredukowane	Przeskalowane	Przeskalowane i Zredukowane
1	MANHC	0.92	0.52	0.94	0.51
2	HC	0.92	0.5	0.92	0.5
3	KMN	0.89	0.52	0.91	0.52
4	KMN++	0.86	0.5	0.89	0.5
5	GMM	0.23	0.58	0.79	0.59
6	GMMK	0.84	0.35	0.58	0.33
7	CMN	0.59	0.32	0.61	0.34
8	KMD	0.6	0.32	0.53	0.33
9	MMM	1	0.54	0	0.18
10	MMMK	0.97	0.54	0.01	0.08
11	MANCMN	0.05	0.13	0.02	0.12

Table 4: Wpisy reprezentują zsumowane sukcesy (poprawnie przypisane klastry), znormalizowane metodą minimum-maksimum w symulowanych mieszaninach wielomianowych

5 Podsumowanie i wnioski

5.1 Zagregowane wyniki

Istnieje mnóstwo metryk. W zależności od przypadku, niektóre mogą być lepsze od innych, zwłaszcza w przypadku niezrównoważonych danych. Jednak tutaj do agregacji użyto najprostszej, przybliżonej liczby sukcesów. Wyniki zostały znormalizowane przy użyciu metody min-max, gdzie najniższy wynik jest ustawiony na 0, najwyższy na 1, a wszystkie inne wyniki są skalowane proporcjonalnie. Wyniki w tabeli reprezentują proporcję prawidłowo przypisanych grup, przy czym wyższy wynik oznacza lepszą efektywność. Na koniec dane zostały posortowane na podstawie największej sumy dla różnych typów danych.

5.2 Symulowane dane

Wyniki badania wykazały, że algorytm GMM wypadł najlepiej w ujęciu ogólnym, uzyskując wynik 1 dla danych kompletnych. Algorytm KMN również wypadł dobrze, z wynikiem od 0,57 do 0,84 w zależności od rodzaju użytych danych. Algorytmy KMN++, GMMK, MANHC i HC miały umiarkowaną wydajność, natomiast algorytmy CMN, MMMK, MMM, MANCMN i KMD miały słabą lub gorszą wydajność. Ogólnie rzecz biorąc, algorytmy GMM i KMN są dobrym wyborem dla klastrowania w tym zbiorze danych, szczególnie przy użyciu kompletnych i skalowanych danych.

Porównując wyniki, algorytm MMM uzyskał najwyższy wynik wśród wszystkich algorytmów., z wynikiem 1. Jednak algorytm MMMK również wypadł dobrze z wynikiem 0,97 na kompletnym zbiorze danych. HC, KMN i KMN++ również wypadły nieznacznie gorzej, z wynikami od 0,86 do 0,92 na pełnym zbiorze danych

Nr.	Algorytm	Kompletne	Zredukowane	Przeskalowane	Przeskalowane i Zredukowane
1	GMM	0.78	0.49	0.51	0.53
2	GMMK	0.76	0.36	0.64	0.5
3	HC	0.42	0.38	0.77	0.51
4	MMMK	0.91	0.82	0.12	0.2
5	KMN++	0.4	0.44	0.66	0.56
6	KMN	0.41	0.43	0.6	0.55
7	MANHC	0.45	0.41	0.64	0.46
8	MMM	1	0.67	0	0.27
9	KMD	0.38	0.39	0.49	0.44
10	CMN	0.12	0.1	0.72	0.32
11	MANCMN	0.15	0.08	0.39	0.31

Table 5: Wpisy reprezentują zsumowane sukcesy (poprawnie przypisane klastry), znormalizowane metodą minimum-maksimum w prawdziwych danych

i podobną wydajnością na pozostałych zbiorach danych. GMMK uzyskał lepsze wyniki niż GMM w przypadku kompletnego zbioru danych, ale gorzej wypadł w przypadku zbiorów zredukowanych, przeskalowanych oraz przeskalowanych i zredukowanych. CMN i KMD miały niższe wyniki w zakresie od 0,32 do 0,61 na kompletnym zbiorze danych, podczas gdy GMMK uzyskał wynik 0,58.

Ogólnie rzecz biorąc, badanie sugeruje, że MMM i MMMK są efektywnymi algorytmami grupowania dla danego rodzaju danych, szczególnie gdy dane są pełnowymiarowe.

5.3 Prawdziwe dane

W badaniu GMM i GMMK miały najwyższe wyniki sumaryczne, przy czym GMM uzyskał najwyższe wyniki w zbiorze danych przeskalowanych, a GMMK w zbiorze danych zredukowanych. Jednak najwyższy wynik należał MMM, gdy dane były kompletne. HC, KMN i KMN++ uzyskały podobne wyniki we wszystkich czterech zestawach danych, podczas gdy MANHC uzyskał nieco lepsze wyniki niż HC. KMD i CMN uzyskały najniższe wyniki spośród wszystkich algorytmów, podczas gdy MANCMN okazał się minimalnie lepszy od KMD i CMN.

Podsumowując, podczas gdy GMM i GMMK miały najwyższe wyniki sumaryczne, należy zauważyć, że MMM osiągnął najwyższy wynik w całym zbiorze danych.

5.4 Wnioski

Celem niniejszego projektu doktorskiego było porównanie wydajności algorytmów dystansowych i modelowych na różnych zbiorach danych. W szczególności w badaniu zaimplementowano dwa algorytmy oparte na modelu (Gaussian Mixture EM i Multinomial Mixture EM). Te dwa algorytmy oraz cztery oparte na odległości (aglomeracyjne grupowanie hierarchiczne, k-means, k-medoids i fuzzy c-means) zostały zastosowane zarówno do symulowanych, jak i rzeczywistych zbiorów danych. W badaniu wykorzystano kilka metryk do oceny wyników grupowania, w tym skorygowany indeks Randa, prosty współczynnik dopasowania, ważony indeks Jaccarda, zrównoważoną dokładność oraz metryki oparte na sprzężonym rozkładzie beta-binomialnym. Metryki te pozwoliły na ilościowe określenie jakości wyników klastrowania oraz porównanie wydajności różnych algorytmów na różnych zbiorach danych. Wyniki badań zostały przedstawione w formie graficznej wraz z krótkim opisem wyników.

Intuicyjnie wiemy, że wydajność różnych algorytmów różni się w zależności od rodzaju i złożoności zbioru danych. Podczas gdy algorytmy oparte na odległości są bardzo potężne, algorytmy oparte na modelu przedstawione w tej pracy są wysoce konkurencyjne i są potężnymi narzędziami w metodach nienadzorowanego grupowania.