© Wydawnictwo UR 2016
ISSN 2080-9069
ISSN 2450-9221 online

„Edukacja – Technika – Informatyka" nr 4/18/2016
www.eti.rzeszow.pl
DOI: 10.15584/eti.2016.4.46

TADEUSZ KWATER[1], **ROBERT PĘKALA**[2], **ALEKSANDRA SALAMON**[3]

Algorytm klasyfikacji obiektów na przykładzie przestrzeni medialnej

The algorithm for the classification of the example of media space

[1] Doktor habilitowany profesor UR, Uniwersytet Rzeszowski, Wydział Matematyczno-Przyrodniczy, Katedra Inżynierii Komputerowej, Polska
[2] Doktor, Państwowa Wyższa Szkoła Techniczno-Ekonomiczna w Jarosławiu, Polska
[3] Studentka, Państwowa Wyższa Szkoła Techniczno-Ekonomiczna w Jarosławiu, Polska

Streszczenie

W artykule zaprezentowano rozwiązanie zagadnienia klasyfikacji obiektów w przestrzeni medialnej. Zastosowano sekwencyjny algorytm grupowania dla wybranych obiektów będących informacjami w portalach internetowych, a reprezentowanych wektorem cech. Uzyskano zadawalające rezultaty klasyfikacji zależne od przyjętego wektora cech i od założonych parametrów wejściowych.

Słowa kluczowe: sekwencyjny algorytm grupowania, nienadzorowana klasyfikacja, przestrzeń medialna, wektor cech.

Abstract

The solution of the problem of classification of objects in the media is presented in the article. Sequential algorithm was used to group the selected objects in selected portals internet. Objects were information's of portals represented by a feature vector. Achieved satisfactory results classification dependent adopted the feature vector and the assumed input parameters.

Key words: the sequence clustering algorithm, unsupervised classification, media space, the feature vector.

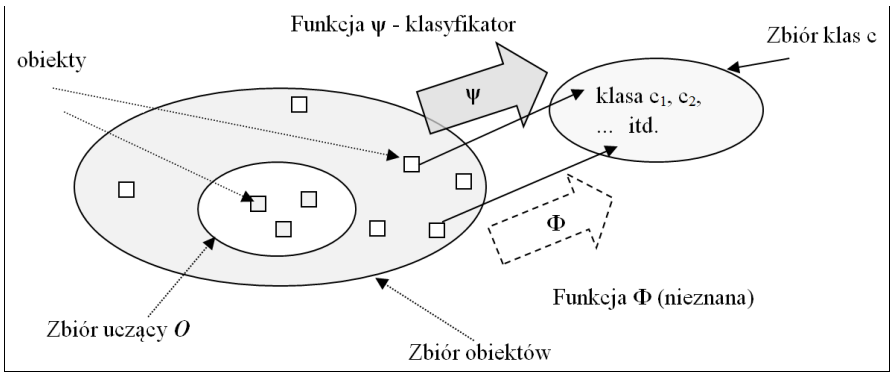
Wstęp

Pojęcie klasyfikacji obiektów jest ściśle związane z ich rozpoznawaniem, czyli dokonaniem podziału na grupy [Stąpor 2011; Klasyfikacja]. Jest to metoda eksploracji danych. Jako obiekt należy rozumieć sygnał, proces czy informację

zwartą, a rozpoznanie jako przyporządkowanie według podobieństw pewnych cech. Klasyfikacja znalazła obecnie zastosowanie wszędzie tam, gdzie wykorzystuję się informatykę. Informacje w przestrzeni medialnej (internet, radio, telewizja) można potraktować jako obiekt scharakteryzowany pewnymi cechami. Wybór cech może być rozmaity i często zależy od rozpoznawanych obiektów. Określenie podobieństwa obiektów oznacza przypisanie ich do zbioru traktowanego jako klasa obiektów. Takie uogólnienie pozwala na odwzorowanie przestrzeni medialnej na przestrzeń z metryczną, w której odległości są mierzone przyjętymi cechami. Podejście związane z klasyfikacją przestrzeni medialnej w takim ujęciu może stanowić pewnego rodzaju system monitorowania; może być pomocne w kreowaniu programów danego portalu, aby stał się atrakcyjnym w sensie zainteresowań odbiorców.

Algorytm grupowania sekwencyjnego

Drążenie danych, pozyskiwanie wiedzy, wydobywanie danych lub ekstrakcja danych zwane klasyfikacją danych są określane jako automatyczne odkrywanie nieznanych wcześniej reguł i zależności w zbiorze danych [Rozpoznawanie; Fatyga, Podraza 2010]. Uzyskane w ten sposób dane mogą być wykorzystane do różnych celów, np. do określania trendów. W przypadku klasyfikacji odpowiedzią jest klasa (kategoria, grupa) obiektu.



Rys. 1. Obraz procesu klasyfikacji obiektów

W celu zunifikowanego podejścia do dokonania odpowiedniego podziału na klasy każdy obiekt jest reprezentowany cechami (mierzonymi lub obliczonymi). Ponadto przyjmuje się, iż proces klasyfikacji (rozpoznawania) jest efektem uprzedniego procesu uczenia z przykładów. W zależności od rodzaju informacji dostarczonej systemowi w zbiorze przykładów uczenie może przebiegać w sposób nadzorowany lub nienadzorowany. W klasyfikacji nadzorowanej zakłada się, że na całej populacji obiektów jest określona pewna (nieznana oczywiście

systemowi) funkcja przypisująca każdemu obiektowi etykietę jednej z klas, a systemowi dostarcza się jedynie mały podzbiór w postaci tzw. zbioru uczącego. System ma za zadanie odnaleźć jak najlepsze (w sensie przyjętego kryterium) przybliżenie **φ** nieznanej mu funkcji **Φ**, które pozwoli przypisywać dowolny obiekt $\boldsymbol{o}_i$ ze zbioru **O** do jednej z $\boldsymbol{c}$ klas. Jednym z wielu składowych odwzorowań **φ** jest funkcja **Ψ**, która odwzorowuje zbiór wszystkich możliwych reprezentacji obiektów w zbiór etykiet klas (zob. rys. 1). Funkcja ta nazywana jest algorytmem klasyfikacji, regułą decyzyjną lub czasami potocznie klasyfikatorem.

W klasyfikatorze danymi wejściowymi są zbiory krotek, zaś danymi wyjściowymi są odpowiedzi, które przydzielają wartość atrybutu każdej krotce. Wartość atrybutu zostaje przydzielona krotce na podstawie wartości pozostałych atrybutów [Rozpoznawanie]. Istnieje wiele metod klasyfikacji obiektów. W artykule zaprezentowano najczęściej stosowaną metodę grupowania sekwencyjnego, zwłaszcza w początkowym etapie klasyfikacji. Jest to metoda nienadzorowana, wykorzystująca pojęcia takie jak wektor cech oraz prototyp, stosowane w klasyfikatorze minimalno-odległościowym. Pojęcie prototyp $\boldsymbol{m_i}$ określono jako środek (średni wektor cech) danej klasy:

$$\mathrm{m_i} = \frac{1}{\mathrm{N_i}} \sum_{i=1}^{N_i} (\mathrm{x}) \quad \mathrm{i}=1, \ldots, \mathrm{c}, \tag{1}$$

gdzie: x – wektor cech obiektu, c – liczba klas, $\mathrm{n_i}$ – liczebność klasy $\mathrm{U_i}$.

Do rozważań przyjęto, iż miarą podobieństwa obiektu reprezentowanego przez wektora cech $\boldsymbol{x_i}$ do grupy $\mathrm{G_k}$ jest odległość euklidesowa $\mathrm{d(x_i, G_k)}$ od prototypu $\boldsymbol{m_i}$ dana w postaci:

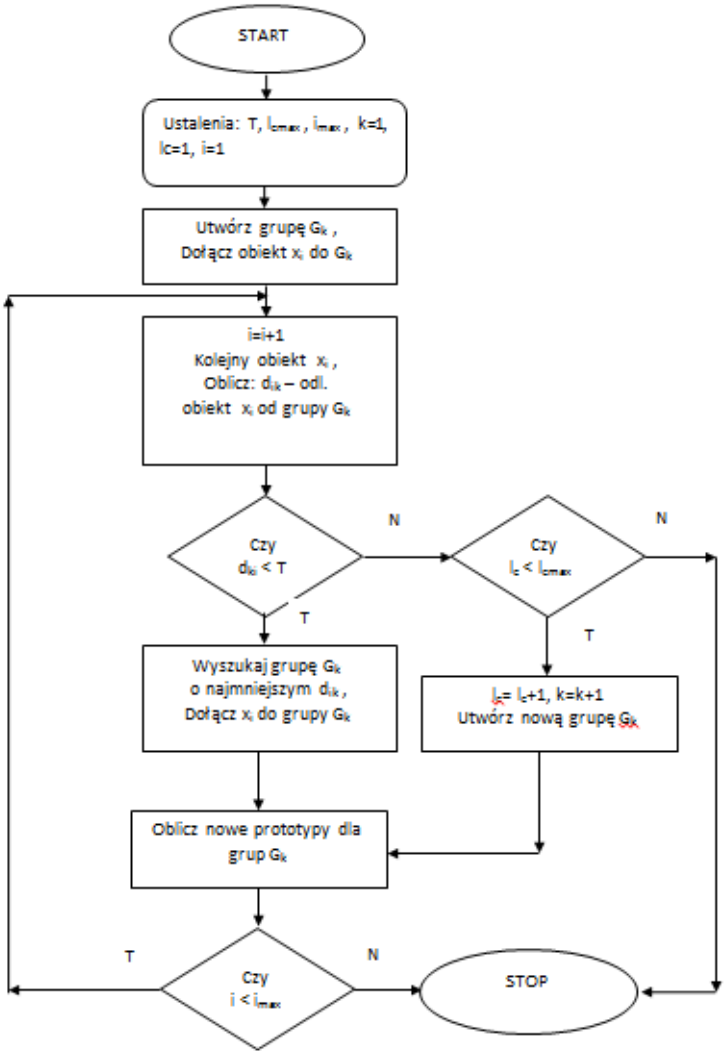
$$\mathrm{d(x_i, G_k)} = \sqrt{(x_i - m_k)^T (x_i - m_k)}\,. \tag{2}$$

Dla tak przyjętych założeń zastosowano sekwencyjny algorytm, którego schemat zamieszczono na rys. 2. Umożliwia on przydzielanie obiektów do odpowiednich klas, jeśli spełnione będą warunki odległościowe (2), przy czym wartość progową tworzenia nowej klasy oznaczono jako $\boldsymbol{T}$, a maksymalną liczbę klas jako $\boldsymbol{l_{cmax}}$.

Eksperymenty, obiekty medialne, specyfikacja cech

Obiektami badań były informacje prezentowane w następujących portalach internetowych: Onet.pl, Wp.pl, Interia.pl, Gazeta.pl, Dziennik.pl. Obejmowały one artykuły z sekcji wiadomości, sport, ekonomia/biznes w lutym 2016 r. Dla sekcji sport została wybrana jedna grupa tematyczna dotycząca startu Justyny Kowalczyk w Pucharze Świata. Natomiast dla sekcji ekonomia/biznes zostały wybrane artykuły/obiekty dotyczące kryzysu w polskich kopalniach. Dla wybranych obiektów jako cechy przyjęto: 1) pozycja w agendzie (W1), 2) gęstość informowania (W2) [Salamon 2016]. Zatem obiekt będzie reprezentowany

2-elementowym wektorem cech. Dla otrzymanych wartości cech zastosowano wstępną obróbkę danych w postaci normalizacji. Badania były przeprowadzane od godziny 8:00 do godziny 18:00.



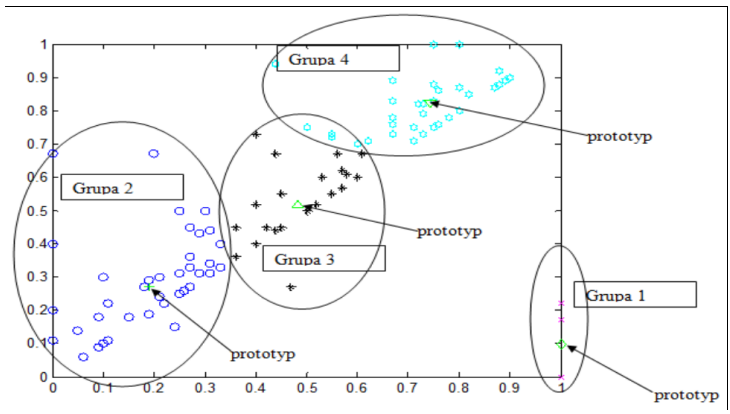
Rys. 2. Schemat algorytmu sekwencyjnego

Sposób wyznaczenia cechy W1 oraz W2 określono z zależności:

$$W1 = \frac{a_1 + a_2 + \ldots + a_n +}{l_{pom}} \quad , \qquad x = \frac{\mathrm{n}}{\mathrm{M} * \mathrm{T}}, \tag{3}$$

gdzie: $a_1, a_2, \ldots, a_n$ – pozycja informacji w danym momencie pomiarowym, l_{pom} – liczba pozycji dla danej informacji, n – liczba pojawień się informacji, M – liczba portali, na których pojawiła się informacja, T – okres czasowy, w ciągu którego gęstość była mierzona.

Po dokonaniu normalizacji współrzędnych wektora cech dla zrealizowanych pomiarów przyjętego progu nieprawdopodobieństwa T = 0,35 i obliczeń zgodnie z algorytmem sekwencyjnym (rys. 2) otrzymano rezultat prezentowany na rys. 3.



Rys. 3. Rezultat klasyfikacji obiektów w przestrzeni medialnej

Podsumowanie

W artykule przedstawiono zagadnienie klasyfikacji obiektów w przestrzeni medialnej. Obiektami były informacje prezentowane w wybranych portalach internetowych. Rozwiązanie dla takiego zagadnienia dokonano, stosując algorytm sekwencyjnego grupowania pozwalającego na przypisanie do poszczególnych grup obiektów o podobnych cechach. Algorytm ten należy do grupy klasyfikacji nienadzorowanej. Interesującym elementem w tym podejściu jest wyznaczenie cech informacji, traktowanej jako obiekt. Przyjęto, iż cechami tymi są: pozycja w agendzie oraz gęstość informacji. Dla tak określonych cech zastosowano sekwencyjny algorytm grupujący obiekty z różnymi miarami podobieństwa traktowanymi jako wartości parametrów wejściowych. W wyniku takiego postępowania uzyskano zadawalające rezultaty końcowe w postaci wyodrębnionych klas obiektów. Otrzymany rozkład klas charakteryzuje się kształtami kulistymi, liczebność klas zależy od przyjętej progowej definicji odległości. Przeprowadzone testy wykazały poprawne funkcjonowanie algorytmu. Dalsze prace badawcze mogłyby obejmować modyfikacje zaprezentowanego algorytmu np. w postaci podwójnej prezentacji wektorów obiektów czy losowy wybór obiektów do klasyfikacji.

Literatura

Fatyga P., Podraza R. (2010), *Klasyfikacja danych – przegląd wybranych metod*, Warszawa.

Klasyfikacja, http://wazniak.mimuw.edu.pl/images/5/5f/ED-4.2-m07-1.0.pdf.

Rozpoznawanie, http://www.eletel.p.lodz.pl/pstrumil/po/rozpoznawanie.pdf.

Salamon A. (2016), *Automatyczna klasyfikacja obiektów na przykładzie przestrzeni medialnej*, praca inż., promotor: T. Kwater, PWSTE, Jarosław.

Stąpor K. (2011), *Automatyczna klasyfikacja obiektów w wizji komputerowej*, Warszawa.