PRZEGLĄD TELEINFORMATYCZNY
NR 1-4, 2021

Rekomendacje jako wynik oceny preferencji na podstawie wskazanych przykładów

Włodzimierz KWIATKOWSKI

Instytut Teleinformatyki i Cyberbezpieczeństwa, Wydział Cybernetyki, WAT,
ul. gen. S. Kaliskiego 2, 00-908 Warszawa
wlodzimierz.kwiatkowski@wat.edu.pl

STRESZCZENIE: Rozpatrywany jest problem wyznaczania rekomendacji na podstawie wskazanych przykładów decyzji akceptowalnych i przykładów decyzji nieakceptowalnych. Wskazanie przez decydenta tych przykładów jest podstawą oceny jego preferencji. Istota przedstawionego rozwiązania polega na określeniu preferencji jako klastra wyznaczonego poprzez uzupełnianie wskazanych przykładów. W artykule zaproponowano procedurę kolejnych przybliżeń bazującą na rozwiązaniach zadania klasyfikacji na podstawie zadanych przykładów.

SŁOWA KLUCZOWE: rekomendacja, preferencje, eksploracja danych, klasyfikacja, grupowanie

1. Wprowadzenie

Rozpatrywane zadanie wyznaczenia rekomendacji polega na sprecyzowaniu, które decyzje z zadanego zbioru są zgodne z preferencjami decydenta. Preferencje decydenta są wyrażane jako zbiór decyzji akceptowalnych przez niego. Sposobem poznania preferencji decydenta jest analiza wskazywanych przez niego przykładów. Wyznaczony na tej podstawie zbiór decyzji akceptowalnych jest traktowany jako rekomendacja dla decydenta.

Definiowanie cech i wskaźników jakości decyzji w sposób niezależny od konkretnego aktu wyboru decydenta praktycznie oznacza, że decydent ma preferencje arbitralnie narzucone. Wnioskowanie o preferencjach decydenta tylko na podstawie wskazywanych przez niego przykładów decyzji ocenionych pozytywnie (akceptowalnych) wyklucza taką sytuację. Takie wnioskowanie oznacza także, że ewaluacja decyzji dokonywana jest bezpośrednio na podstawie

ich charakterystyk (np. pomiarów, obserwacji), a nie na podstawie narzuconych cech i wskaźników jakościowych.

Sformułowanie problemu wyznaczania rekomendacji na podstawie wskazanych przez decydenta przykładów decyzji jest przedstawione w [3]. Przyjęta tam metoda ewaluacji opiera się na wyznaczaniu w przestrzeni cech odległości analizowanych decyzji od decyzji wskazanych.

Omawiane w niniejszym artykule zadanie różni się od przedstawionego w [3]. Różnica wynika z innej interpretacji wskazywanych przykładów. W [3] wskazane przez decydenta przykłady są traktowane jako deklaracja jego preferencji. W konsekwencji tego podane przykłady stanowią ustalony zbiór wzorców. W niniejszym artykule wskazywane decyzje są interpretowane jako deklaracja niepełna. Istota przedstawianego w artykule rozwiązania polega na określeniu preferencji decydenta jako klastra (skupienia) wyznaczonego poprzez uzupełnianie wskazanych przykładów.

W przypadku istnienia sprzężenia zwrotnego wzbogacanie zbioru przykładów polega na cyklicznym zapoznawaniu decydenta z proponowaną rekomendacją i uzyskiwaniu od niego dodatkowych wskazówek. Istota metody proponowanej w niniejszym artykule polega na wykorzystaniu idei stopniowego wzbogacania zbioru decyzji wskazanych przez decydenta. Przyjmuje się przy tym założenie, że proces modyfikacji powinien następować bez udziału decydenta. Podstawą możliwości realizacji tej idei jest analiza wskazywanych przykładów na tle wszystkich rozpatrywanych decyzji. Taką analizę umożliwia przedstawiona w [4] metoda regularyzacji zadań klasyfikacji.

Rozpatrywany w artykule problem jest sformułowany jako poszukiwanie metody wyznaczania rekomendacji na podstawie wskazania przez decydenta przykładów decyzji akceptowalnych, a także przykładów decyzji nieakceptowalnych. Celem wskazywania przykładów decyzji nieakceptowalnych jest racjonalne ograniczanie liczebności rekomendowanych decyzji.

2. Prace związane

Sformułowane zadanie wyznaczania rekomendacji można zaliczyć do projektowania systemów eksploracji danych, których wyróżnikiem jest wyszukiwanie informacji według zgłoszonego zapotrzebowania użytkownika. Przyjęte założenia powodują, że poszukiwane jest rozwiązanie polegające na filtrowaniu decyzji opartym zarówno na treści (ang. *content based filtering*), jak i na współpracy (ang. *collaborative filtering*) [6].

Podstawowym przykładem iteracyjnego grupowania poprzez kolejne modyfikacje wyników jest algorytm ISODATA (ang. *Iterative Self-Organizing Data Analysis Techniques*) [1]. Określenie *algorytm ISODATA* jest najczęściej

rozumiane jako szczególna metoda nienadzorowanej klasyfikacji [2]. Obecnie algorytm ISODATA jest wykorzystywany np. przy klasyfikacji obrazów multi-spektralnych[1].

Zasadniczy kłopot pojawiający się przy klasyfikacji na podstawie wskazywanych przykładów wynika z faktu, że wskazywane przykłady generują podprzestrzeń, której wymiar jest mniejszy od wymiaru przestrzeni cech. Problem wynikający z faktu, że liczba wskazanych przykładów wzorców jest mała względem liczby współrzędnych wektora cech, jest rozpatrywany w [5]. Zaproponowane są tam dwie metody optymalizacji bazujące na wyznaczaniu rzutów wektorów cech na podprzestrzenie wzorców. Wyróżnikiem pierwszej metody jest wykorzystywanie odległości wektora cech od podprzestrzeni wzorców. Druga metoda polega na przeniesieniu zadania optymalizacji do podprzestrzeni wzorców. Przedstawiona w [4] metoda regularyzacji jest rozwiązaniem kompromisowym i pozwala efektywnie wykorzystywać wskazania decydenta także w przypadkach osobliwych macierzy kowariancji cech wskazanych przykładów.

3. Klasyfikacja na podstawie zadanych wzorców klas

Dany jest zbiór decyzji ponumerowany od 1 do N. Dla każdej decyzji znany jest jej wektor cech. Dla decyzji o numerze k stosować będziemy następujące oznaczenie wektora cech:

$$\mathbf{a}_k = \left[a_{1,k},\quad a_{2,k},\quad \ldots\ ,\quad a_{L,k}\right]^T, \qquad \mathbf{a}_k \in R^L \tag{1}$$

Każda współrzędna $a_{l,k}$ jest liczbą rzeczywistą, a parametr L określa liczbę współrzędnych wektora cech. Wektory cech zadanego zbioru decyzji zestawiamy w postaci następującej macierzy:

$$\mathbf{A} = \left[\mathbf{a}_1,\quad \mathbf{a}_2,\quad \ldots\ ,\quad \mathbf{a}_N\right], \qquad \mathbf{a}_k \in R^L \tag{2}$$

Macierz kowariancji wektorów cech wyznaczana jest następująco:

$$\mathbf{R} = \tfrac{1}{N-1}\sum_{k=1}^{N}\left(\mathbf{a}_k - \overline{\mathbf{a}}\right)\left(\mathbf{a}_k - \overline{\mathbf{a}}\right)^T \tag{3}$$

gdzie:

$$\overline{\mathbf{a}} = \tfrac{1}{N}\sum_{k=1}^{N}\mathbf{a}_k \tag{4}$$

[1] Na podstawie: https://en.wikipedia.org/wiki/Multispectral_pattern_recognition. Dostęp: 23.08.2021.

Przyjmiemy dalej, że macierz kowariancji wektorów cech jest nieosobliwa:

$$\det(\mathbf{R}) \neq 0 \tag{5}$$

Odległość pomiędzy wektorami $\mathbf{x}$, $\mathbf{y}$ przestrzeni cech R^L będziemy wyznaczać w sposób uwzględniający wielkość rozrzutu (rozproszenia) wartości współrzędnych oraz ich wzajemną korelację. Wymagania te spełnia odległość Mahalanobisa, jest ona określona wzorem:

$$d_e(\mathbf{x},\mathbf{y}) = \sqrt{(\mathbf{x}-\mathbf{y})^T \mathbf{R}^{-1}(\mathbf{x}-\mathbf{y})}, \qquad \mathbf{x},\mathbf{y} \in R^L \tag{6}$$

Wskazania przykładów określających wzorzec klasy o indeksie $h \in \{1,2,\ldots,H\}$ (gdzie: H – liczba klas) będziemy dokonywać przez podanie odpowiedniego zbioru indeksów W_h. Liczbę elementów wzorca o indeksie h oznaczamy jako

$$N_h = \|W_h\| \tag{7}$$

Wzorzec klasy o indeksie h jest reprezentowany przez następujący zbiór punktów (klaster) w przestrzeni cech:

$$\mathrm{C}(W_h) = \{\mathbf{a}_k \in R^L : k \in W_h\} \tag{8}$$

Wnioskowanie o podobieństwie cechy $\mathbf{x}$ do wzorca o indeksie h bazuje na określeniu odległości $D_e(\mathbf{x}, \mathrm{C}(W_h))$ punktu $\mathbf{x}$ od klastra $\mathrm{C}(W_h)$. Przykładowo, wybierając metodę centroidalną wyznaczania odległości między klastrami, otrzymujemy zależność:

$$D_e(\mathbf{x},\mathrm{C}(W_h)) = d_e(\mathbf{x},\overline{\mathbf{w}}_h) = \sqrt{(\mathbf{x}-\overline{\mathbf{w}}_h)^T \mathbf{R}^{-1}(\mathbf{x}-\overline{\mathbf{w}}_h)} \tag{9}$$

gdzie:

$$\overline{\mathbf{w}}_h = \frac{1}{N_h} \sum_{j \in W_h} \mathbf{a}_j \tag{10}$$

Klasyfikacja oparta na wykorzystywaniu metryki (6) nazywana jest środowiskową [4].

Stosowanie klasyfikacji środowiskowej znajduje uzasadnienie wtedy, gdy cechy wszystkich wzorców są jednorodne w następującym sensie: odpowiednie klastry różnią się wartościami oczekiwanymi, a odpowiadające im macierze kowariancji są jednakowe. W przypadku, gdy macierze kowariancji wzorców różnią się, zalecane jest zróżnicowanie sposobu pomiaru odległości stosownie do macierzy kowariancji poszczególnych wzorców [2].

Macierz kowariancji wyznaczoną na podstawie przykładów wzorca o indeksie h oznaczymy następująco:

$$\mathbf{R}_h = \frac{1}{N_h - 1} \sum_{j \in W_h} \left(\mathbf{a}_j - \overline{\mathbf{w}}_h\right)\left(\mathbf{a}_j - \overline{\mathbf{w}}_h\right)^T \tag{11}$$

Odległość pomiędzy wektorami $\mathbf{x}$, $\mathbf{y}$ przestrzeni cech R^L zadaną wzorem:

$$d_h(\mathbf{x},\mathbf{y}) = \sqrt{(\mathbf{x}-\mathbf{y})^T \mathbf{R}_h^{-1} (\mathbf{x}-\mathbf{y})}\,, \qquad \mathbf{x},\mathbf{y} \in R^L \tag{12}$$

nazywa się dopasowaną do wzorca o indeksie h [2]. Podobnie nazywać będziemy odległość między cechą $\mathbf{x}$ a klastrem $\mathrm{C}(W_h)$. Przykładowo dla centroidalnej metody grupowania odległość ta jest określona wzorem:

$$D_h(\mathbf{x},\mathrm{C}(W_h)) = d_h(\mathbf{x},\overline{\mathbf{w}}_h) = \sqrt{(\mathbf{x}-\overline{\mathbf{w}}_h)^T \mathbf{R}_h^{-1} (\mathbf{x}-\overline{\mathbf{w}}_h)} \tag{13}$$

Klasyfikacja względem zadanych wzorców polega na przyporządkowaniu analizowanej decyzji o indeksie k do klasy o indeksie h wtedy, jeśli [2]

$$D_h(\mathbf{a}_k,\mathrm{C}(W_h)) = \min_{j=1,2,\ldots,H} D_j(\mathbf{a}_k,\mathrm{C}(W_j)) \tag{14}$$

Potrzeba regularyzacji występuje, gdy macierze kowariancji cech wzorców są osobliwe lub źle uwarunkowane (występuje duża rozpiętość między ich wartościami własnymi, a ich wyznaczniki są bliskie zeru). Zaproponowana w [4] metoda regularyzacji polega wykorzystaniu następującej metryki:

$$d_{h,\rho}(\mathbf{x},\mathbf{y}) = \sqrt{(\mathbf{x}-\mathbf{y})^T \mathbf{R}_{h,\rho}^{-1} (\mathbf{x}-\mathbf{y})}\,, \qquad \mathbf{x},\mathbf{y} \in R^L \tag{15}$$

przy czym:

$$\mathbf{R}_{h,\rho} = (1-\rho)\mathbf{R}_h + \rho\mathbf{R} \tag{16}$$

gdzie: $\rho \in [0,1]$ – współczynnik regularyzacji. Metryka $d_{h,\rho}(\mathbf{x},\mathbf{y})$ różni się od metryki $d_h(\mathbf{x},\mathbf{y})$ zastąpieniem macierzy kowariancji $\mathbf{R}_h$ kombinacją wypukłą tej macierzy i macierzy kowariancji $\mathbf{R}$. Wartość $\rho = 0$ oznacza brak regularyzacji i klasyfikację dopasowaną, wartość $\rho = 1$ oznacza przejście do klasyfikacji środowiskowej.

4. Rekomendacje na podstawie klasyfikacji względem wzorców

Podstawowe założenie poczynione przy wyznaczaniu przedstawianego w tym punkcie rozwiązania jest następujące: wskazane przykłady decyzji akceptowalnych są traktowane jako wzorzec decyzji akceptowalnych, a wskazane przykłady decyzji nieakceptowalnych – jako wzorcowe dla odpowiadającej im klasy. Możliwe jest zastosowanie dwóch metod klasyfikacji: środowiskowej

i dopasowanej do wzorców [3]. W obu przypadkach rozwiązanie zadania klasyfikacji można bezpośrednio wykorzystać do wyznaczania rekomendacji. Uzyskiwane rozwiązanie przedstawia podział zbioru decyzji na dwie rozdzielne klasy: decyzji akceptowalnych i decyzji nieakceptowalnych. Wyznaczoną klasę decyzji akceptowalnych można traktować jako rekomendację dla decydenta. Metoda ta obarczona jest wrodzoną wadą: rekomendowane są wszystkie decyzje, którym jest bliżej do przykładów akceptowalnych niż do przykładów nieakceptowalnych. Można tę wadę osłabić, zadając odpowiednio małą liczebność zbioru rekomendowanych decyzji lub nakładając na decyzje rekomendowane jakościowe ograniczenia. Takie ograniczenia liczebności zbioru decyzji rekomendowanych są możliwe na podstawie uszeregowania elementów wyznaczonej klasy na podstawie ich odległości od swojego wzorca.

Jest oczywiste, że rekomendacje można wyznaczać zarówno dla decyzji akceptowalnych, jak i nieakceptowalnych. Jest to niewątpliwa zaleta wykorzystywania zadania klasyfikacji. Podział zbioru decyzji na dwie przeciwstawne klasy jest najbardziej ostrym opisem preferencji decydenta.

5. Rekomendacje na podstawie szeregowania decyzji

Podobnie jak poprzednio, wskazane przykłady decyzji akceptowalnych są traktowane jako wzorce decyzji akceptowalnych. Podstawą wyznaczenia rekomendacji jest uporządkowanie (uszeregowanie) zbioru decyzji według odległości decyzji od klastra wzorców, rozpoczynając od decyzji położonej w odległości najmniejszej. Do zbioru decyzji rekomendowanych kwalifikowane są kolejne decyzje zgodnie z tym uszeregowaniem. Jest oczywiste, że wskazane przykłady decyzji nieakceptowalnych należy wykluczyć.

Bardziej wnikliwe wnioskowanie o wykluczaniu decyzji w procesie kwalifikowania do rekomendacji jest możliwe na podstawie uporządkowania (uszeregowania) zbioru decyzji względem wzorca decyzji nieakceptowalnych. W tym przypadku należy więc wykonać uporządkowanie zbioru wszystkich decyzji zarówno względem wzorców decyzji akceptowalnych, jak i uporządkowanie tego zbioru względem wzorców decyzji nieakceptowalnych. Konfrontacja tych dwóch przeciwstawnych uszeregowań powinna umożliwić racjonalne ograniczanie liczebności zarówno zbioru rekomendowanych decyzji akceptowalnych, jak i zbioru rekomendowanych decyzji nieakceptowalnych. Proponowana idea ograniczania sekwencji jest następująca. Po napotkaniu w szeregu decyzji uporządkowanych według wzorców akceptowalnych decyzji wskazanej jako przykład (wzorzec) decyzji nieakceptowalnej z procesu rekomendowania wyklucza się także wszystkie decyzje następne w szeregu (usytuowane dalej od klastra wzorców decyzji akceptowalnych). Analogicznie można ograniczać rekomendacje decyzji nieakceptowalnych.

Realizacja przedstawionej wyżej idei polega na niezależnym przeprowadzeniu dwóch procesów kwalifikowania decyzji: do zbioru decyzji rekomendowanych jako akceptowalne i do zbioru decyzji rekomendowanych jako nieakceptowalne. W obu przypadkach proces kwalifikowania do odpowiedniej rekomendacji należy zakończyć po napotkaniu w szeregu decyzji wskazanej jako kontrprzykład. Po wykluczeniu elementów wspólnych uzyskane klastry mogą być traktowane jako odpowiednie, przeciwstawne rekomendacje.

6. Metoda iteracyjna wyznaczania rekomendacji na podstawie szeregowania

Wskazanie wzorcowych przykładów przez decydenta pośrednio daje informację o tym, które współrzędne wektora cech i jakie ich wartości, są dla decydenta istotne.

Poszerzenie wymiaru podprzestrzeni generowanej przez wzorce można uzyskać poprzez zwiększenie liczebności zbioru wzorców. Proponujemy w tym celu, aby rekomendacje uzyskane na podstawie wskazanych przez decydenta przykładów zinterpretować jako nowy zbiór decyzji definiujących odpowiednią klasę decyzji: akceptowalnych bądź nieakceptowalnych. Uznawanie rekomendacji za nowy wzorzec klasy można powtarzać.

Ideę proponowanej metody można przedstawić jako iteracyjne uzupełnianie klastra wzorców decyzji akceptowalnych i klastra wzorców decyzji nieakceptowanych. Uzupełnianie polega na wyznaczaniu w każdej iteracji klastra decyzji rekomendowanych jako akceptowalne i klastra decyzji rekomendowanych jako nieakceptowalne. W kolejnej iteracji wyznaczone klastry decyzji rekomendowanych stają się odpowiednimi klastrami wzorców. Efektywną metodą wyznaczania przeciwległych rekomendacji jest opisana wcześniej metoda bazująca na podwójnym szeregowaniu decyzji: względem wzorców decyzji akceptowalnych oraz względem wzorców decyzji nieakceptowalnych. Istotną dla procesu rekomendowania jest eliminacja elementów wspólnych w obu sekwencjach. Procedurę tę można uzupełnić ograniczeniami liczebności zbiorów decyzji rekomendowanych oraz ograniczeniami natury jakościowej. Efektywnym działaniem zapewniającym te efekty jest ograniczanie maksymalnej odległości analizowanej decyzji od klastra wzorcowego.

Oczekiwanym rezultatem opisanego, iteracyjnego procesu jest uzyskanie ustalonych zbiorów rekomendowanych decyzji. Zbiory te nie powinny ulegać zmianie w kolejnych iteracjach. Uzyskane, ustalone klastry (skupienia) stanowią bezpośredni opis preferencji decydenta, zarówno akceptowalności, jak i nieakceptowalności decyzji. Bezpośredni opis preferencji oznacza tu wyliczenie

(enumerację) wszystkich preferowanych decyzji. Utożsamienie tak rozumianego opisu preferencji z odpowiednią rekomendacją jest naturalną konsekwencją.

7. Eksperyment obliczeniowy

7.1. Przedmiot i cel badań

Celem badań było eksperymentalne potwierdzenie uzyskiwania użytecznych rekomendacji poprzez iteracyjne uzupełniania wskazanych przez decydenta przeciwstawnych przykładów. Istotnym badanym problemem była ocena procesu uzyskiwania rekomendacji, a w szczególności potwierdzenie uzyskiwania rekomendacji ustalonych (tj. nie zmieniających się w kolejnych iteracjach).

Użyte w badaniach dane pomiarowe zostały pobrane z archiwalnych baz danych stacji IMGW[2]. W wybranym do analizy zbiorze pomiarów „decyzja" oznacza pojedynczą stację i jest scharakteryzowana wektorem $L = 2$ pomiarów. Analizowana baza danych zawiera wyniki pomiarów 62 stacji. Pomiary zostały wykonane w tym samym dniu.

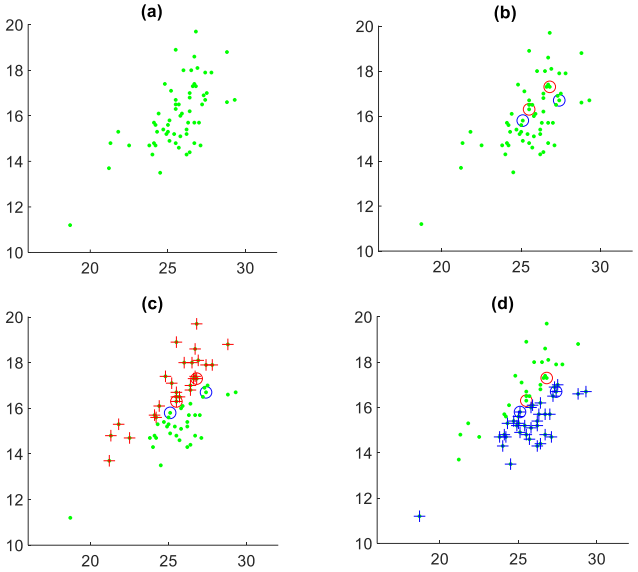
Przyjęty do obliczeń wymiar przestrzeni cech zapewnia czytelną wizualizację wyników. Szczególnie ilustracja przypadku małej liczby wskazanych przykładów powinna być łatwo interpretowalna (wskazanie tylko dwóch przykładów sugeruje ograniczenie preferencji do punktów leżących na prostej przechodzącej przez wektory cech wskazanych przykładów).

7.2. Wyniki wyznaczania rekomendacji środowiskowych

Wyniki wyznaczania rekomendacji na podstawie klasyfikacji środowiskowej stanowią punkt wyjścia do oceny proponowanej w artykule metody.

Źródłem danych jest macierz złożona z 62 wektorów pomiarów wykonanych przez poszczególne stacje. Wektory te określają środowisko eksperymentu. Dla przyjętych do obliczeń danych macierz kowariancji wektorów pomiarowych jest nieosobliwa. Umożliwia to oparcie obliczeń na metryce zdefiniowanej wzorem (6).

[2] Źródło danych*: IMGW Dane pomiarowo-obserwacyjne. Dane meteorologiczne, dobowe, klimatyczne.* Plik: k_d_07_2020.csv (2020_07_k.zip). URL: https://dane.imgw.pl/data/dane_pomiarowo_obserwacyjne/dane_meteorologiczne/dobowe/klimat/2020/. Dostęp: 23.08.2021.

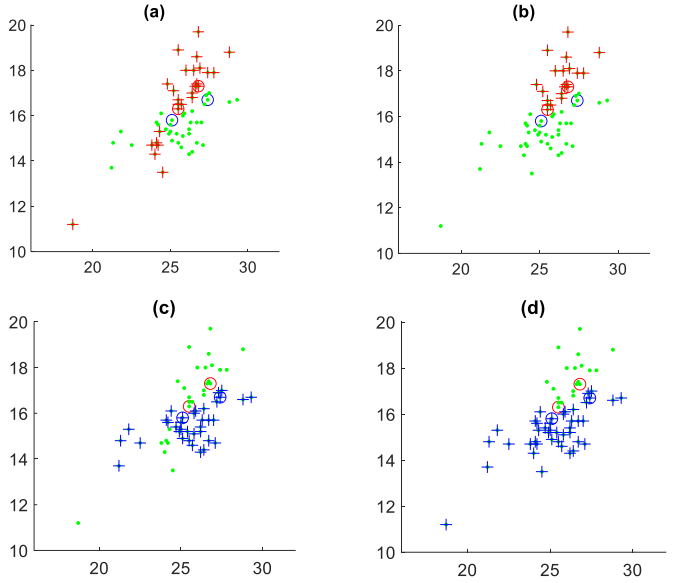


Rys. 1. Wizualizacja wyznaczania rekomendacji środowiskowych. (a) Decyzje środowiska w przestrzeni cech. (b) Wskazane przykłady (wzorce). (c) Rekomendacje dla klasy A (decyzji akceptowalnych). (d) Rekomendacje dla klasy NA (decyzji nieakceptowalnych). Punkty zielone oznaczają wektory cech decyzji środowiska. Wektory cech wskazanych (wzorcowych) decyzji oznaczono kółkami: dla klasy A kolorem czerwonym, dla klasy NA kolorem niebieskim. Wektory cech decyzji rekomendowanych oznaczono znakiem plus w odpowiednim kolorze

Klasyfikacja została przeprowadzona na podstawie wskazanych przykładów: klasy A (decyzji akceptowalnych) oraz wskazanych przykładów klasy NA (decyzji nieakceptowalnych). Przyjęto, że liczba zarówno przykładów klasy A, jak i przykładów klasy NA jest równa 2. Założenie to uniemożliwia bezpośrednie wykorzystywanie odległości dopasowanych do poszczególnych klas (macierze kowariancji dla wskazywanych przykładów klas są osobliwe). Można zauważyć, że w ogólnym przypadku przestrzeni cech R^L do wykonania klasyfikacji dopasowanej potrzebne jest wskazanie co najmniej $L+1$ przykładów (w dwuwymiarowej przestrzeni cech pożądane jest wskazanie co trzech przykładów). W praktyce dla dużych wartości L często okazuje się to istotnym problemem, zwłaszcza w przypadku uzyskiwania przykładów w wyniku współpracy z decydentem-człowiekiem.

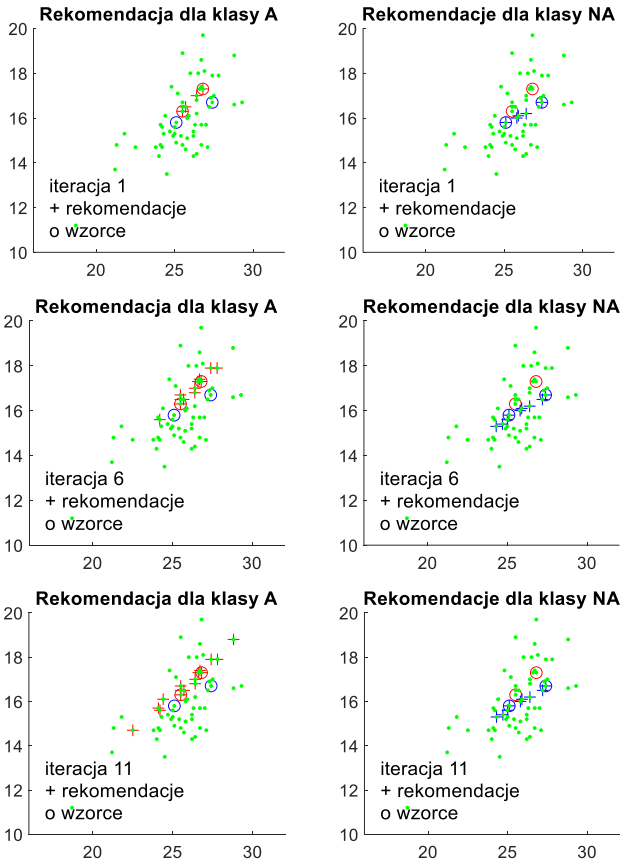
Obliczenia wykonano dla klasyfikacji środowiskowej oraz – po zastosowaniu regularyzacji – dla klasyfikacji dopasowanej do wzorców. Wyniki klasyfikacji środowiskowej są przedstawione na rys.1. Przyjęte do obliczeń przykłady wzorców zostały specjalnie dobrane dla potrzeb wizualizacji.

Przedstawione zobrazowanie ukazuje przypadek, kiedy każda analizowana decyzja jest kwalifikowana jako rekomendacja decyzji akceptowalnej bądź nieakceptowalnej. Mimo oczywistego podejrzenia, że decydent, podając swoje, nieliczne przykłady, raczej nie określił tak szeroko swoich preferencji.



Rys. 2. Wizualizacja wyznaczania rekomendacji dopasowanych na podstawie zadania regularyzowanego. (a) Rekomendacje dla klasy A wyznaczone dla współczynnika regularyzacji $\rho=10^{-10}$. (b) Rekomendacje dla klasy A wyznaczone dla współczynnika regularyzacji $\rho=10^{-1}$. (c) Rekomendacje dla klasy NA wyznaczone dla współczynnika regularyzacji $\rho=10^{-10}$. (d) Rekomendacje dla klasy NA wyznaczone dla współczynnika regularyzacji $\rho=10^{-1}$. Punkty zielone oznaczają wektory cech decyzji środowiska. Wektory cech wskazanych (wzorcowych) decyzji oznaczono kółkami: dla klasy A kolorem czerwonym, dla klasy NA kolorem niebieskim. Wektory cech decyzji rekomendowanych oznaczono znakiem plus w odpowiednim kolorze

Na rys. 2 przedstawiono wyniki obliczeń dla klasyfikacji dopasowanej do wzorców. Przyjęta do obliczeń liczba wskazywanych przykładów (dwa dla każdej klasy) powoduje konieczność wykorzystywania regularyzacji zadania. Wskazane wektory cech decyzji akceptowalnych i ich kontrprzykłady są położone blisko siebie, a proste generowane przez odpowiednie punkty przecinają się, tak aby pokazać wpływ współczynnika regularyzacji na uzyskiwane wyniki. Także w tym przypadku każda analizowana decyzja jest kwalifikowana jako rekomendacja decyzji akceptowalnej bądź nieakceptowalnej.



Rys. 3. Wizualizacja wyznaczania rekomendacji metodą iteracyjną na podstawie regularyzowanego zadania klasyfikacji. Punkty zielone oznaczają wektory cech decyzji środowiska. Wektory cech wskazanych (wzorcowych) decyzji oznaczono kółkami: dla klasy A kolorem czerwonym, dla klasy NA kolorem niebieskim. Wektory cech decyzji rekomendowanych oznaczono znakiem plus w odpowiednim kolorze

7.3. Wyniki iteracyjnego wyznaczania rekomendacji metodą iteracyjnego uzupełniania wskazanych przykładów

W odróżnieniu od poprzednio przedstawianych obliczeń wskazania decydenta nie są traktowane jako zamknięte listy wzorców klas, a jedynie ich przykłady. Podstawą zakwalifikowania do decyzji rekomendowanych jest szeregowanie decyzji względem wskazanych przykładów. Proces kwalifikowania kolejnych decyzji jest kończony w momencie, gdy kolejna, analizowana decyzja jest już zakwalifikowana do klasy przeciwstawnej. Oczekiwanym wynikiem

procesu kwalifikacji jest wzbogacenie wyjściowych przykładów obu przeciwstawnych klas. Wynikową rekomendacją jest zbiór decyzji nieulegający zmianom w kolejnej iteracji.

Na rys. 3 przedstawiono wyniki wyznaczania rekomendacji metodą iteracyjnego uzupełniania wskazanych przez decydenta przykładów decyzji akceptowalnych oraz nieakceptowalnych. Do obliczeń przyjęto takie same wskazania jak w obliczeniach poprzednich. Przyjęto wartość współczynnika regularyzacji $\rho = 0{,}005$. Ograniczenie liczebności klastrów było uzyskiwane przez ustalenie maksymalnej, dopasowanej do wzorca odległości decyzji od średniej wartości klastra wzorca. Brak zmian rozwiązania zaobserwowano po jedenastu iteracjach.

8. Podsumowanie

Przedstawione sformułowanie problemu bazuje na zadaniu wyznaczania rekomendacji w procesie decyzyjnego wspomagania decydenta w wyszukiwaniu informacji w dużych bazach danych (np. w diagnostyce medycznej, automatycznym poszerzaniu zbioru uczącego w zadaniach uczenia sieci neuronowych). Takie ujęcie problemu ułatwia interpretację algorytmu, w tym dobór uniwersalnego słownictwa. Obszar zastosowań można ogólniej określić jako analizę danych z niestandardowego punktu widzenia.

Zasadniczy wynik przedstawionych w artykule propozycji stanowi konstatacja, że współpracę z decydentem przy wyznaczaniu ograniczonych rekomendacji można zredukować do jednorazowego wskazania przykładów i kontrprzykładów.

Przedstawiane rozwiązanie bazuje na interpretacji dokonanych przez decydenta wskazań jako przykładowych wzorców dwóch przeciwstawnych klas. Istota proponowanej metody polega na równoległym szeregowaniu decyzji względem wzorców każdej klasy. Rekomendowane klasy uzyskuje się przez wzajemne ograniczanie rekomendowanych sekwencji przez wzorce przeciwstawne. Uzyskane w ten sposób rekomendacje stają się przykładowymi wzorcami w następnej iteracji.

Wymiar generowanej przez przykłady podprzestrzeni cech jest ograniczony przez liczbę wskazywanych przykładów. Wynikający stąd problem polega na możliwości odrzucania decyzji, których wektory cech nie leżą w wygenerowanej przez przykłady podprzestrzeni. Narzędziem pomagającym eliminować takie przypadki jest regularyzacja. Stosowanie zbyt dużych wartości współczynnika regularyzacji prowadzi jednak do rekomendacji środowiskowej.

Przedstawione w artykule wyniki uzyskano przy wykorzystywaniu metody centroidalnej obliczania odległości między klastrami. Większą wrażliwość na

zmiany pojedynczych decyzji można uzyskać stosując metody bardziej złożone [2].

Ocena preferencji decydenta na podstawie porównywania odległości między klastrami prowadzi do grupowania cech o wartościach skorelowanych. Użyteczność tej metody jest widoczna zwłaszcza w przypadku dysponowania obserwacjami (cechami) niefiltrowanymi z punktu widzenia preferencji decydenta (tzn. kiedy poszczególne cechy nie są wyróżnikami preferencji).

Satysfakcja decydenta-człowieka z uzyskanej rekomendacji jest trudna do przewidzenia. Ulotność problemu jest konsekwencją niewiedzy decydenta. Jest skutkiem ograniczonej zdolności decydenta do ewaluacji dużej liczby decyzji. W zadaniach automatycznego przeszukiwania baz danych (np. w celu poszukiwania podobnych obiektów) można formułować ocenę jakościową proponowanej procedury.

Uzyskanie rozwiązań trywialnych (np. pustych klastrów) może być skutkiem niespójności dokonanych przez decydenta wskazań, wynikających bądź z niedopasowania przestrzeni obserwacji (cech) do sygnalizowanych preferencji decydenta, bądź ze sprzecznych jego wskazań.

Literatura

[1] BALL G.H., HALL D.J., *Isodata, an Iterative Method of Multivariate Analysis and Pattern Classification.* Proceedings of the IFIPS Congress, 1965.

[2] KWIATKOWSKI W., *Metody automatycznego rozpoznawania wzorców*. BEL Studio, Warszawa, 2010.

[3] KWIATKOWSKI W., *Recommendations as a result of decision evaluations based on reference examples*, Teleinformatics Review, No. 1-2, 2019, pp. 3-23.

[4] KWIATKOWSKI W., *The regularization method in the classification task according to given examples* Teleinformatics Review, No. 3-13, 2019, pp. 3-23.

[5] KWIATKOWSKI W., *Wykrywanie anomalii bazujące na wskazanych przykładach.* Przegląd Teleinformatyczny, nr 1-2, 2018, s. 3-21.

[6] MOBASHER B., DAI H., LUO T., NAKAGAWA M., *Improving the Effectiveness of Collaborative Filtering on Anonymous Web Usage Data*. In: Proceedings of the IJCAI 2001, Workshop on Intelligent Techniques for Web Personalization (ITWP01), 2001.

Recommendations as a result of the assessment of preferences on the basis of the indicated examples

ABSTRACT: The problem of determining a decision recommendation according to examples of acceptable decisions and examples of unacceptable decisions indicated by the decision-maker is considered in the paper. The decision-maker's examples are the foundation for assessing his preferences. The essence of the presented solution consists in determining the preferences of the decision-maker as a cluster designated by supplementing the indicated examples. The paper proposes a procedure of successive approximations based on the classification task according to given examples.

KEYWORDS: recommendation, preferences, data mining, classification, clustering

Praca wpłynęła do redakcji: 18.10.2021 r.