

# STATYSTYCZNE METODY ROZPOZNAWANIA OBRAZÓW I ICH ZASTOSOWANIA

Małgorzata Misztal

*Katedra Metod Statystycznych, Uniwersytet Łódzki, Łódź*

## 1 WPROWADZENIE

Szybko zmieniające się warunki stosowania określonych metod analizy statystycznej czy metod modelowania ekonometrycznego do rozwiązywania problemów o charakterze decyzyjnym wymuszają zmianę podejścia do dotychczas wykorzystywanych procedur badawczych i metod diagnostyki czy predykcji statystycznej.

Obserwowany w ostatnich latach rozwój techniki komputerowej wywarł ogromny wpływ na powstanie nowych dziedzin nauki, które wymagają zapamiętania i przetwarzania dużej ilości danych opisanych w przestrzeniach wielowymiarowych w celu efektywnego rozwiązywania praktycznych problemów.

Złożoność algorytmów i czas obliczeń przesłały już stanowić barierę rozwoju narzędzi usprawniających szeroko pojętą działalność człowieka, w tym także metod wspomagania procesów podejmowania decyzji. Nowe techniki zbierania informacji statystycznej, wymuszone przez komputerowe bazy danych, skłaniają do stosowania metod umożliwiających opracowanie i przeanalizowanie informacji w możliwie krótkim czasie i dla nieskończonej dużej zbiorów danych statystycznych.

Takie właśnie metody proponuje teoria rozpoznawania obrazów, przy czym obraz rozumiany jest jako ilościowy opis obiektu, zdarzenia czy zjawiska.

Ogólnie zadanie teorii rozpoznawania obrazów polega na określaniu przynależności rozmaitego typu obiektów do pewnych klas. Rozpoznawanie to przebiega w sytuacji braku apriorycznej informacji co do reguł przynależności obiektów do poszczególnych klas,

a jedyną dostępną informację stanowi zwykle tzw. ciąg uczący, złożony z obiektów, których prawidłową klasyfikację znamy (tzw. rozpoznawanie z nauczycielem).

Najczęściej wykorzystywane, teoriodecyzyjne metody rozpoznawania wymagają przyjęcia założenia, że rozpoznawany obiekt, scharakteryzowany wartościami  $p$  cech, może być rozpatrywany jako punkt  $\mathbf{x}=(x_1, \dots, x_p)^T$   $p$ -wymiarowej przestrzeni  $\mathbf{X}$  ( $\mathbf{X} \subseteq \mathbb{R}^n$ ) i traktowany jako realizacja wektora losowego  $X$  o funkcji gęstości  $f_i(\mathbf{x})$ ,  $i \in \mathbf{K}$  ( $\mathbf{K}=\{1, \dots, k\}$  - jest zbiorem numerów klas).

Algorytmem rozpoznawania  $\psi$  (algorytmem klasyfikacji, regułą decyzyjną) nazywamy przepis, według którego odbywa się przyporządkowanie rozpoznawanemu obiektowi  $\mathbf{x} \in \mathbf{X}$  numeru klasy  $i \in \mathbf{K}$ :  $\psi(\mathbf{x}) = i$ . Innymi słowy, mamy tu do czynienia z odwzorowaniem przestrzeni cech w zbiór numerów klas:  $\psi: \mathbf{X} \rightarrow \mathbf{K}$ , bądź też z generowaniem rozkładu przestrzeni cech na rozłączne obszary decyzyjne:  $R_i = \{\mathbf{x} \in \mathbf{X}: \psi(\mathbf{x})=i\}$ ,  $i \in \mathbf{K}$  [8].

W rozpoznawaniu teoriodecyzyjnym do opisu sytuacji wykorzystuje się modele probabilistyczne i statystyczne, ze względu na ich szczególną przydatność do wykrywania niepewnych i niejednoznacznych związków między klasami i ilościowymi charakterystykami obiektów.

## 2 CELE I ZAŁOŻENIA PRACY

Zasadniczym celem pracy jest ocena wybranych klasycznych i nieklasycznych metod rozpoznawania obrazów.

Cele szczegółowe określone są następująco:

- Prezentacja i klasyfikacja wybranych algorytmów rozpoznawania obrazów w zależności od *a priori* posiadanej informacji na temat rozkładów prawdopodobieństwa charakteryzujących losowy związek między klasami i cechami.
- Ocena właściwości (w sensie dokładności predykcji) klasycznych algorytmów rozpoznawania na podstawie badań eksperymentalnych.
- Stworzenie "rankingu" algorytmów rozpoznawania poprzez określenie zasad wyboru z dużej liczby algorytmów rozpoznawania metody najlepszej (w sensie prostoty konstrukcji, precyzji klasyfikacji i łatwości implementacji komputerowej).
- Ocena skuteczności klasycznych i nieklasycznych metod rozpoznawania z punktu widzenia efektywności decyzji w konkretnych problemach badawczych.

Teza rozprawy została ujęta następująco: drzewa klasyfikacyjne mogą być traktowane jako uniwersalne narzędzie tworzenia reguł przynależności obiektów do klas.

### 3 METODY TWORZENIA ALGORYTMÓW ROZPOZNAWANIA

Rozpoznawanie obrazów można zdefiniować jako wieloetapowy proces przetwarzania informacji, podczas którego relatywnie duża ilość danych wejściowych zostaje przetworzona na mniejszą ilość danych użytecznych, zakończony klasyfikacją czyli przypisaniem obiektowi numeru klasy [1].

Wśród metod tworzenia algorytmów rozpoznawania wyróżniamy podejście oparte na modelu probabilistycznym oraz podejście oparte na modelu statystycznym.

W przypadku modelu probabilistycznego zakłada się, że dla każdego rozpoznawanego obiektu  $\mathbf{x}$  znane jest prawdopodobieństwo *a priori* zdarzenia, że pochodzi on z klasy  $i$ -tym numerze, a także znane są warunkowe gęstości rozkładów cech w poszczególnych klasach:

$$f(\mathbf{x}/i) = f_i(\mathbf{x}).$$

W takiej sytuacji możliwe jest obliczenie wskaźnika jakości rozpoznawania oraz, poprzez rozwiązanie odpowiedniego problemu optymalizacyjnego, wyznaczenie reguły decyzyjnej

minimalizującej ten wskaźnik. W zadaniach rozpoznawania opartych na modelach probabilistycznych wykorzystuje się np. klasyfikację bayesowską lub regułę minimaxową (por. np. [7], [8], [11]).

Model statystyczny jest z kolei podstawą konstrukcji reguł decyzyjnych ze zbiorem uczącym, złożonym z obiektów, dla których znany jest wektor wartości cech oraz numer klasy. Wśród metod rozpoznawania ze zbiorem uczącym rozważać można dwie sytuacje:

1 Znamy z założenia postać funkcyjną warunkowych gęstości w klasach a nie znamy ich parametrów – dokonujemy zatem ich estymacji na podstawie zbioru uczącego. Wśród algorytmów rozpoznawania opartych na parametrycznym modelu statystycznym szczególną uwagę zwraca się na te metody, w których przyjmuje się założenie o normalności rozkładów cech obiektów w klasach. W tym przypadku przedstawić można zwykłe, bayesowskie i quasi-bayesowskie estymatory kwadratowych i liniowych funkcji klasyfikujących i dyskryminujących a także algorytmy wykorzystujące odległości Rao i Mahalanobisa (por. np. [7]).

2 Brak jest jakichkolwiek założeń co do postaci funkcyjnej warunkowych gęstości w klasach – dokonuje się więc estymacji funkcji gęstości za pomocą metod nieparametrycznych. W tym celu wykorzystać można np. algorytm rozpoznawania oparty na estymatorze Parzena oraz algorytm oparty na przedziałach zmienności cech. Szczególnym przypadkiem nieparametrycznych metod rozpoznawania są algorytmy minimalnoodległościowe, bazujące na pojęciach sąsiedztwa i odległości. Wśród minimalnoodległościowych algorytmów rozpoznawania wymienić należy algorytm najbliższego sąsiada (NN),  $\alpha$  najbliższych sąsiadów ( $\alpha$ -NN) oraz  $\alpha$ -tego najbliższego sąsiada ( $\alpha$ -th NN) a także algorytm oparty na odległościach między obiektami (por. np. [5], [8], [14]).

Dodatkowo, wspomnieć warto także o metodzie Mojirsheibaniego [10] tworzenia algorytmów kombinowanych, będących złożeniem kilku reguł decyzyjnych.

Wymienione algorytmy rozpoznawania określić można mianem klasycznych, bazują one bowiem na rozwiązaniach analizy dyskryminacji, metod decyzji statystycznych, teorii

estymacji (zarówno parametrycznej jak i nieparametrycznej), bayesowskiej teorii decyzji lub metod optymalizacyjnych.

Alternatywę dla omówionych metod rozpoznawania stanowią nieklasyczne metody określania reguł przynależności obiektów do klas. Szczególną uwagę zwrócić tu należy na analizę drzew klasyfikacyjnych, bowiem metoda rekurencyjnego podziału jest stosunkowo mało znana, zwłaszcza w zastosowaniach ekonomiczno – społecznych.

Metoda rekurencyjnego podziału polega na stopniowym podziale  $p$ -wymiarowej przestrzeni cech na rozłączne podzbiory aż do uzyskania ich homogeniczności ze względu na wyróżnioną cechę. W wyniku rekurencyjnego podziału zbiór uczący  $U$  zostaje podzielony na  $M$  rozłącznych podzbiorów  $U_1, U_2, \dots, U_M$  zgodnie z następującą procedurą (por. [4]):

- 1 Dla danego zbioru obiektów sprawdzić, czy jest on jednorodny ze względu na wartości zmiennej zależnej lub spełnione jest inne, przyjęte kryterium stopu. Jeśli tak – zakończyć postępowanie.
- 2 Jeśli nie – rozważyć wszystkie możliwe podziały zbioru  $U$  na rozłączne podzbiory  $U_1, U_2, \dots, U_M$  w oparciu o wartości kolejno wybieranych zmiennych objaśniających.
- 3 Ocenic jakość każdego z podziałów zgodnie z przyjętym kryterium i wybrać najlepszy z nich.
- 4 Podzielić zbiór obiektów w wybrany sposób. Kroki 1-4 wykonać rekurencyjnie dla każdego podzbioru  $U_1, U_2, \dots, U_M$ .

Procedurę podziału kończymy, jeżeli zostało osiągnięte założone kryterium stopu – zwykle jednorodność podzbiorów  $U_1, U_2, \dots, U_M$  lub określona, minimalna liczebność podzbiorów. Proces rekurencyjnego podziału zbioru  $U$  można przedstawić graficznie w postaci drzewa klasyfikacyjnego.

Wśród algorytmów tworzących drzewa klasyfikacyjne wymienić można m. in. CART (Classification and Regression Trees [2]), QUEST (Quick, Unbiased, Efficient Statistical Trees [9]), CRUISE (Classification Rule with Unbiased Interaction Selection and Estimation [6]).

Do nieklasycznych metod rozpoznawania zaliczyć można także sieci neuronowe, powstałe na gruncie teorii biocybernetycznej.

## 4 OCENA JAKOŚCI WYBRANYCH ALGORYTMÓW ROZPOZNAWANIA

Wyniki przedstawione w pracy obejmują:

- Ocenę wybranych klasycznych metod rozpoznawania;
- Sformułowanie reguł ułatwiających decydentowi wybór spośród dużej liczby algorytmów rozpoznawania metod najlepszych w celu efektywnego rozwiązania praktycznych zadań;
- Zastosowanie i ocenę efektywności metod rozpoznawania obrazów przy wspomaganie procesów podejmowania decyzji w rzeczywistych problemach badawczych.

Porównanie wybranych klasycznych algorytmów rozpoznawania wymagało przeprowadzenia eksperymentu Monte Carlo.

### 4.1 Założenia eksperymentu Monte Carlo

Rozpoznawaniu podlegały obiekty należące do dwóch klas ( $k=2$ ). W praktycznych zastosowaniach algorytmów rozpoznawania klasyfikacja obiektów należących do dwóch klas ma miejsce najczęściej. Np. w diagnostyce medycznej rozważamy osoby, u których wystąpiła bądź nie dana jednostka chorobowa; w analizach finansowych bankowcy dzielą przedsiębiorstwa na takie, które są w stanie spłacić zaciągnięte kredyty i te, którym kredytu udzielać nie należy; w badaniach demograficznych rozważać można podział województw na te, w których występuje ujemny przyrost naturalny i takie, w których przyrost jest dodatni, itp.

Przyjęto, że każdy obiekt opisany jest czterowymiarowym wektorem obserwacji  $\mathbf{x}=[x_1, x_2, x_3, x_4]^T$ . Wartości każdej cechy wektora obserwacji generowano z rozkładów jednowymiarowych, zgodnie z założonymi parametrami.

Rozpatrzono trzy warianty rozkładów (I– klasy najbardziej zbliżone; III– klasy najbardziej oddalone). Do konstrukcji algorytmu rozpoznawania wykorzystywano zbiór uczący zaś jakość uzyskanej reguły klasyfikacyjnej oceniano na podstawie klasyfikacji obiektów zbioru testowego.

W konkretnych sytuacjach badawczych zjawiskiem bardzo częstym jest posiadanie mało licznego zbioru uczącego. Wiąże się to zwykle z szeroko rozumianymi kosztami pozyskiwania informacji, np. w analizach medycznych zbioru

o niewielkiej liczebności stanowić mogą pacjenci poddani badaniom obciążającym lub zmarli podczas operacji; w analizach finansowych, z kolei, banki mogą zasłaniać się tajemnicą bankową i nie udostępniać danych dotyczących niewypłacalnych kredytobiorców, itd. Dlatego też rozpatrzono dwa warianty liczebności zbioru uczącego:

- $N=20$  obiektów ( $N_1=7$ ;  $N_2=13$ );
- $N=60$  obiektów ( $N_1=27$ ;  $N_2=33$ ).

W obu przypadkach liczba obiektów ciągu testowego była taka sama:  $n=30$  obiektów ( $n_1=15$ ;  $n_2=15$ ).

Ocenie poddane zostały następujące algorytmy rozpoznawania:

- 1 Algorytm najbliższego sąsiada z miarami odległości Euklidesa, Czekanowskiego, Canberra, Jeffreysa i Matusity.
- 2 Algorytm  $\alpha$  najbliższych sąsiadów z miarami odległości jak w przypadku algorytmu najbliższego sąsiada dla liczby sąsiadów równej 3 oraz 5.
- 3 Algorytm  $\alpha$ -tego najbliższego sąsiada z miarami odległości oraz liczbą sąsiadów jak w przypadku algorytmu  $\alpha$  najbliższych sąsiadów.
- 4 Algorytm rozpoznawania wykorzystujący odległość Mahalanobisa.
- 5 Algorytmy rozpoznawania wykorzystujące zwykłe obciążone, zwykłe nieobciążone, bayesowskie i quasi-bayesowskie estymatory liniowych funkcji klasyfikujących.
- 6 Algorytmy rozpoznawania wykorzystujące zwykłe obciążone, zwykłe nieobciążone, bayesowskie i quasi-bayesowskie estymatory liniowych funkcji kwadratowych.
- 7 Algorytm oparty na odległościach wykorzystujący funkcje klasyfikacyjne z miarami odległości Euklidesa, Czekanowskiego i Canberra.
- 8 Algorytm oparty na estymatorze Parzena z gaussowską funkcją jądra.

Każdy eksperyment powtórzono 1000 razy dla każdego z sześciu rozpatrywanych wariantów (3 warianty rozkładu \* 2 warianty liczebności).

Porównywalność wyników zapewniło wykorzystanie tych samych prób w każdym z rozważanych algorytmów. Obliczenia wykonano w pakiecie GAUSS.

## 4.2 Podstawowe wyniki

Analiza wyników uzyskanych w przeprowadzonym eksperymencie dla poszczególnych wariantów rozkładu i liczebności pozwala na sformułowanie następujących wniosków:

- 1 Zwiększenie liczby obiektów ciągu uczącego prowadzi do zmniejszenia odsetków błędnych klasyfikacji niezależnie od rozważanego algorytmu rozpoznawania.
- 2 Wśród algorytmów minimalnoodległościowych najgorsze rezultaty klasyfikacji dostajemy dla algorytmu  $\alpha$ -tego najbliższego sąsiada. Dodatkowo – zwiększenie wartości  $\alpha$  w tym przypadku prowadzi do wzrostu odsetka błędnych klasyfikacji.
- 3 Algorytm oparty na odległościach pozwala poprawić dokładność klasyfikacji w stosunku do algorytmów NN,  $\alpha$ -NN i  $\alpha$ -tego-NN dla odległości Euklidesa oraz Czekanowskiego w przypadku klas, których środki ciężkości leżą niedaleko od siebie.
- 4 Liniowe funkcje klasyfikacyjne dają niskie odsetki błędnych klasyfikacji niezależnie od typu estymatora.
- 5 Kwadratowe funkcje klasyfikujące dla małej liczby obiektów ciągu uczącego ( $N=20$  obiektów) dają istotnie gorsze rezultaty w przypadku estymatorów zwykłego nieobciążonego i estymatora bayesowskiego. Po zwiększeniu liczebności próby uczącej ( $N=60$  obiektów) nie ma różnic w odsetkach błędnych rozpoznań między rozważanymi estymatorami kwadratowych funkcji klasyfikacyjnych.
- 6 Dla niewielkiej liczebności ciągu uczącego liniowe funkcje klasyfikacyjne dają mniejsze odsetki błędnych rozpoznań niż funkcje kwadratowe. Dla bardziej licznych prób zależność ta jest odwrotna.
- 7 Algorytm z estymatorem jądrowym Parzena najlepsze wyniki daje w przypadku najbardziej oddalonych klas (wariant III rozkładu) – są to najmniejsze odsetki błędnych klasyfikacji w stosunku do wszystkich pozostałych rozważanych reguł klasyfikacyjnych.

Przeprowadzona analiza ułatwia sformułowanie pewnych reguł, które mogą być następnie wykorzystane w praktyce.

- 1 W przypadku prób uczących o niewielkiej liczebności użyteczną metodą klasyfikacji są funkcje liniowe, odporne na odstępstwa od normalności rozkładu cech w klasach. Zwykłe estymatory funkcji liniowych są także

łatwo dostępne w pakietach statystycznych (np. *STATISTICA*). Przy większej liczbie elementów ciągu uczącego poprawę jakości klasyfikacji dają funkcje kwadratowe (np. zwykły estymator funkcji kwadratowych), gdyż zachodzi wówczas asymptotyczna normalność.

- 2 Wśród algorytmów rozpoznawania bazujących na odległościach nie poleca się stosować algorytmu  $\alpha$ -tego najbliższego sąsiada, dla którego dostajemy istotnie gorsze oszacowania prawdopodobieństw błędnej klasyfikacji. Algorytmy najbliższego sąsiada i  $\alpha$ -najbliższych sąsiadów dają porównywalne wyniki. Na podkreślenie zasługuje fakt, iż macierze odległości wg formuły np. Euklidesa czy miejskiej można w prosty sposób obliczyć korzystając z ogólnie dostępnych pakietów statystycznych. Poprawę rezultatów klasyfikacji daje algorytm oparty na odległościach.
- 3 Pewną wadą metod minimalnoodległościowych jest to, że wymagają przechowywania całego ciągu uczącego, bowiem klasyfikacja każdego nowego obiektu wymaga obliczenia jego odległości od wszystkich obiektów ciągu uczącego, co znacznie wydłuża czas obliczeń. Przykładowo dla wariantu Ia obliczenia w przypadku algorytmu 3-NN zajęły około 5 minut, w przypadku algorytmu opartego na odległościach – 4 minuty a w przypadku liniowych funkcji klasyfikacyjnych (4 estymatory) – 30 sekund. Dla wariantu Ib czas pracy komputera wyniósł odpowiednio: 19 minut, 17 minut i 40 sekund. Trudno także jednoznacznie wskazać najlepszą miarę odległości. Wybór miary odbywać się może tylko na drodze eksperymentalnej – z kilku czy kilkunastu sprawdzonych miar wybieramy tę, dla której dostajemy niższe odsetki błędnych klasyfikacji.
- 4 W zastosowaniach praktycznych metod rozpoznawania problemem staje się wybór algorytmu dla obiektów opisanych zestawem cech mieszanych. Zwrócić trzeba uwagę na fakt, że usunięcie z wektora obserwacji zmiennych jakościowych znacznie zubaża analizę. Stąd w przypadku liniowych i kwadratowych funkcji klasyfikujących oraz odległości Mahalanobisa konieczna staje się transformacja zmiennych jakościowych na wektory zmiennych zerojedynkowych. Prowadzi to do zwiększenia wymiaru przestrzeni,

co jest zjawiskiem niekorzystnym w przypadku niewielkiej liczebności ciągu uczącego. Z drugiej strony, metody te są mało kosztowne ze względu na czas obliczeń – klasyfikacja obiektów próby testowej wymaga przechowywania w pamięci jedynie współczynników funkcji klasyfikujących.

- 5 Trudności występują także przy wyborze miary odległości dla zmiennych mieszanych do algorytmów minimalnoodległościowych. Pewnym rozwiązaniem jest np. miara Gowera czy odległość kombinowana Cessie i Houwelingen [3].

Zasadne wydaje się zatem stwierdzenie, że wybór algorytmu rozpoznawania zależy od postawionego zadania. W sytuacjach rzeczywistych klasyczne metody rozpoznawania wymagają modyfikacji stosownych do rozważanego, konkretnego problemu, chociaż przeprowadzone badania symulacyjne mogą być podstawą do rekomendacji niektórych reguł klasyfikacyjnych.

## 5 ZASTOSOWANIA ALGORYTMÓW ROZPOZNAWANIA OBRAZÓW

Przedstawione algorytmy rozpoznawania ze zbiorem uczącym znajdują zastosowanie w wielu konkretnych problemach badawczych z różnych dziedzin nauki.

Zaprezentujemy propozycje wykorzystania niektórych metod rozpoznawania w procesie podejmowania decyzji w szeroko rozumianych naukach przyrodniczych i ekonomiczno-społecznych. Przy wyborze metod rozpoznawania kierować się można kryterium użyteczności praktycznej, prostoty interpretacji wyników i dostępności programów realizujących algorytmy rozpoznawania.

Wybrane algorytmy rozpoznawania wykorzystano do klasyfikacji obiektów z 9 zbiorów danych (dane rzeczywiste). Każdy z analizowanych zbiorów w sposób losowy dzielono na zbiór uczący i zbiór testowy. Jako miernik jakości algorytmu przyjęto odsetek błędnych klasyfikacji w próbie testowej (dla bardziej licznych zbiorów danych) lub oszacowanie błędu klasyfikacji metodą leave-one-out.

Do obliczeń wykorzystano:

- Pakiet *STATISTICA PL* – moduły: Analiza dyskryminacyjna oraz Drzewa klasyfikacyjne (algorytmy CART i QUEST);

- Własne programy napisane w *STATISTICA Basic* realizujące algorytmy najbliższego sąsiada,  $\alpha$  najbliższych sąsiadów, algorytm oparty na odległościach z miarami odległości Euklidesa, Czekanowskiego, Canberra, mieszaną, algorytm wykorzystujący odległość Mahalanobisa, algorytmy wykorzystujące zwykle estymatory kwadratowych funkcji klasyfikujących oraz zwykle, bayesowskie i quasi-bayesowskie estymatory liniowych funkcji klasyfikacyjnych;
- Udostępnione w Internecie przez autorów wersje programów tworzących drzewa klasyfikacyjne – QUEST i CRUISE.

Dla zwiększenia przejrzystości prowadzonych analiz wyodrębniono cztery grupy przykładów z różnych dziedzin nauki:

#### 1 Diagnostyka medyczna:

- klasyfikacja pacjentów poddanych PTCA (2 klasy, 6 cech, w tym 2 jakościowe);
- klasyfikacja pacjentów po przeszczepie szpiku (2 klasy, 5 cech, w tym 3 jakościowe);
- klasyfikacja pacjentów z miokardiopatią (3 klasy, 5 cech);
- klasyfikacja pacjentów poddanych CABG (2 klasy, 13 cech, w tym 7 jakościowych).

#### 2 Badania histologiczne:

- klasyfikacja świnek morskich ze względu na poziom amin katecholowych (3 klasy, 5 cech).

#### 3 Problemy społeczno-ekonomiczne:

- analiza ryzyka kredytowego (2 klasy, 10 cech, w tym 5 jakościowych);
- klasyfikacja przedsiębiorstw ze względu na osobę menedżera (3 klasy, 5 cech, w tym 1 binarna);
- klasyfikacja przedsiębiorstw opisanych za pomocą wskaźników ekonomicznych (4 klasy, 5 cech).

#### 4 Psychologia:

- charakterystyka przyczyn narkomanii wśród młodzieży (2 klasy, 4 cechy).

Przeanalizujemy dla przykładu zadanie rozpoznawania, w którym obiektami podlegającymi klasyfikacji są osoby w wieku licealnym, zagrożone narkomanią.

Wyróżnione zostały dwie klasy obiektów: klasa 1 – NIE – osoby nie zażywające środków odurzających; klasa 2 – TAK – osoby narkotyzujące się.

Każda osoba została opisana czterowymiarowym wektorem cech. Zmienne te obrazują

liczbę punktów uzyskanych w testach psychologicznych:

- PZR – poczucie zrozumiałości;
- PZ – poczucie zaradności;
- PS – poczucie sensowności;
- WW – poczucie własnej wartości.

Zbiór danych podzielono losowo na próbe uczącą i testową o liczebnościach:  $NU_1=20$ ,  $NU_2=20$ ;  $NT_1=10$ ,  $NT_2=10$ . Wyniki klasyfikacji przedstawia tabela 1 oraz rysunki 1 i 2.

Tabela 1. Błędne klasyfikacje dla zbioru osób zagrożonych narkomanią.

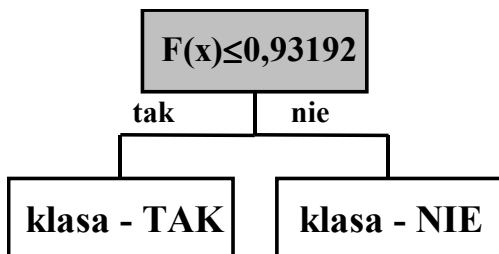
Algorytm rozpoznawania		Błędne klasyfikacje dla zbioru testowego			
NN	O. Czekanowskiego	7/20			
	O. Euklidesa	6/20			
	O. Canberra	7/20			
	O. Jeffreysa–Matusity	7/20			
$\alpha$ -NN	odległości	3-NN	5-NN	7-NN	
	Czekanowskiego	8/20	6/20	6/20	
	Euklidesa	7/20	8/20	6/20	
	Canberra	6/20	7/20	7/20	
	Jeffreysa–Matusity	7/20	8/20	5/20	
$\alpha^{\text{th}}$ -NN	odległości	3 <sup>th</sup> -NN	5 <sup>th</sup> -NN	7 <sup>th</sup> -NN	
	Czekanowskiego	7/20	6/20	12/20	
	Euklidesa	8/20	11/20	12/20	
	Canberra	6/20	9/20	10/20	
	Jeffreysa–Matusity	7/20	9/20	11/20	
Odległości Mahalanobisa		3/20			
Liniowe funkcje klasyfikacyjne		(1) $\hat{e}_1(x)$	(2) $\hat{e}_1(x)$	(3) $\hat{e}_1(x)$	(4) $\hat{e}_1(x)$
		3/20	3/20	3/20	3/20
Kwadratowe funkcje klasyfikacyjne		(1) $\hat{u}_1(x)$	(2) $\hat{u}_1(x)$	(3) $\hat{u}_1(x)$	(4) $\hat{u}_1(x)$
		5/20	5/20	5/20	5/20
Algorytm oparty na odległościach	Euklidesa	4/20			
	Czekanowskiego	4/20			
	Canberra	5/20			
Metody nieparametryczne - estymator Parzena		h=0,40	h=0,55	h=0,70	
		6/20	6/20	6/20	
QUEST szacowane prawdopodobieństwa <i>a priori</i> ; drzewo wielowymiarowe – podział w oparciu o kombinacje liniowe; reguła stopu – 1SE; uzyskane drzewo (rys. 1) ma dwa węzły końcowe;		3/20			
CRUISE szacowane prawdopodobieństwa <i>a priori</i> ; podziały jednowymiarowe; reguła stopu – 1SE uzyskane drzewo (rys. 2) ma trzy węzły końcowe;		2/20			

Reguła klasyfikacyjna uzyskana w wyniku zastosowania algorytmu QUEST (por. rys. 1) jest następująca:

Obiekt  $x$  klasyfikujemy do klasy TAK – narkomani – jeśli wartość funkcji dyskryminacyjnej  $F(x) \leq 0$ ; w przeciwnym wypadku rozpoznawany obiekt zaliczamy do grupy osób nie biorących narkotyków.

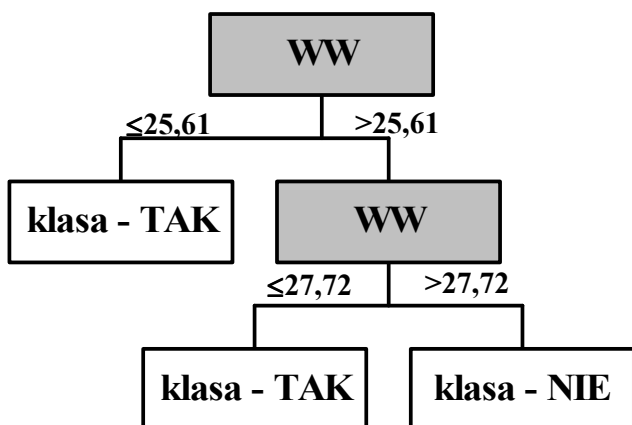
Funkcja dyskryminacyjna w węźle ma postać:

$$F(x) = -0,93192 - 0,0063 * PZR - 0,0174 * PZ + 0,0074 * PS + 0,0602 * WW.$$



Rys. 1. Drzewo klasyfikacyjne dla zbioru osób zagrożonych narkomanią – algorytm QUEST.

Drzewo klasyfikacyjne uzyskane w wyniku zastosowania algorytmu CRUISE przedstawia rys. 2.



Rys. 2. Drzewo klasyfikacyjne dla zbioru osób zagrożonych narkomanią – algorytm CRUISE.

Reguła klasyfikacyjna uzyskana po zastosowaniu algorytmu CRUISE brzmi następująco: osoby o poczuciu własnej wartości wyższym od 27,72 pkt. klasyfikujemy do grupy nie biorących narkotyków; osoby o niskim poczuciu własnej wartości – do 27,72 pkt. – do grupy narkomanów.

Jak łatwo zauważyć, najgorsze klasyfikacje otrzymujemy dla algorytmu  $\alpha$ -tego najbliższego sąsiada.

Wykorzystanie algorytmu z odległością Mahalanobisa oraz liniowych funkcji klasyfikacyjnych (we wszystkich czterech wariantach)

daje niski odsetek błędnych klasyfikacji (po 3 obiekty) mimo braku spełnienia założeń o wielowymiarowej normalności (co sprawdzono uogólnionym testem normalności Shapiro-Wilka).

Taką samą liczbę błędnych zaklasyfikowań daje algorytm QUEST z wielowymiarowymi podziałami za pomocą kombinacji liniowych.

Najlepszy rezultat otrzymujemy dla drzewa utworzonego za pomocą algorytmu CRUISE (2 obiekty błędnie zaklasyfikowane). Zauważmy, iż do podziałów wykorzystywana jest tutaj tylko jedna z cech – poczucie własnej wartości. Zatem klasyfikacja nowej osoby wymaga podania jej tylko testowi określającemu poczucie własnej wartości.

## 6 UWAGI KOŃCOWE

Szczegółowa analiza wyników uzyskanych podczas rozwiązywania realnych zadań rozpoznawania pozwala sformułować wniosek, że w praktycznych zastosowaniach niegorsze (a zwykle lepsze) wyniki klasyfikacji (najmniejszy błąd klasyfikacji szacowany na podstawie zbioru testowego lub sprawdzania krzyżowego) dają algorytmy tworzące drzewa klasyfikacyjne.

Zwrócić należy uwagę na fakt, że procedury tworzenia drzew klasyfikacyjnych nie mają wymagań co do rozkładu badanych zmiennych i są odporne na obserwacje nietypowe.

Drzewa klasyfikacyjne nie stawiają warunków dotyczących pomiaru badanych zmiennych a także umożliwiają klasyfikację obrazów opisanych wektorem cech z wartościami brakującymi. Uzyskane w wyniku analizy drzew klasyfikacyjnych reguły decyzyjne są proste w interpretacji a klasyfikacja obiektów ciągu testowego nie wymaga zwykle pomiaru wszystkich cech objaśniających, co zmniejsza koszty prowadzonych analiz.

Wymienione zalety i dostępność oprogramowania pozwalają uznać metody tworzenia drzew klasyfikacyjnych za użyteczne i precyzyjne narzędzie rozpoznawania, alternatywne w stosunku do metod klasycznych.

Przedstawione przykłady zastosowań metod rozpoznawania wskazują, że mogą one być szeroko wykorzystywane do wspomagania procesów podejmowania decyzji w każdym aspekcie działalności człowieka.

Omówione algorytmy rozpoznawania nie wyczerpują oczywiście problematyki konstrukcji reguł klasyfikacyjnych. Dalsze kierunki badań obejmować będą te metody rozpoznawania, w których obok ciągu uczącego wykorzystuje się zbiór reguł ekspertów, stanowiący w tym przypadku komplementarny sposób pozyskiwania wiedzy na potrzeby algorytmu rozpoznawania.

Zadanie rozpoznawania, w którym zakładamy jednoczesną znajomość zbioru uczącego i reguł eksperta ma duże walory praktyczne, gdyż w rzeczywistych przykładach zastosowań metod rozpoznawania oba rodzaje danych są uzupełniającymi się źródłami informacji o odmiennym pochodzeniu i uzyskanymi w różny sposób.

Wspólne rozpatrzenie dwóch jakościowo różnych typów danych, w inny sposób ujmujących związki między klasami i cechami, powinno być inspiracją nowych idei zmierzających do jednoczesnego wykorzystania tych danych w algorytmie rozpoznawania.

- 13) StatSoft, Inc. (1997). *STATISTICA PL dla Windows (Tom V): Języki: STATISTICA BASIC i SCL*. Kraków: StatSoft Polska.
- 14) Tadeusiewicz, R., Flasiński, M. (1991). *Rozpoznawanie obrazów*. Warszawa: PWN.

## BIBLIOGRAFIA

- 1) Bobrowski, L. (1987). *Dyskryminacja symetryczna w rozpoznawaniu obrazów. Teoria, algorytmy, zastosowania w komputerowym wspomaganie diagnostyki medycznej*. Wrocław: Ossolineum.
- 2) Breiman, L., Friedman, J., Olshen, R., Stone, C. (1984). *Classification and Regression Trees*. London: CRC Press.
- 3) Cessie, S., Houwelingen, H. C. (1995). Testing the Fit of a Regression Model via Score Tests in Random Effects Models. *Biometrics*. 1995, Vol. 51, No 2, pp. 600-614.
- 4) Gatnar, E. (2001). *Nieparametryczna metoda dyskryminacji i regresji*. Warszawa: PWN.
- 5) Jajuga, K. (1990). *Statystyczna teoria rozpoznawania obrazów*. Warszawa: PWN.
- 6) Kim, H., Loh, W.-Y. (2001). Classification Trees with Unbiased Multiway Splits, *Journal of the American Statistical Association*. 2001, Vol. 96, pp. 598-604.
- 7) Krzyśko, M. (1990). *Analiza dyskryminacyjna*. Warszawa: WNT.
- 8) Kurzyński, M. (1997). *Rozpoznawanie obiektów. Metody statystyczne*. Wrocław: Oficyna Wydawnicza Politechniki Wrocławskiej.
- 9) Loh, W.-Y., Shih, Y.-S. (1997). Split Selection Methods for Classification Trees. *Statistica Sinica*. 1997, Vol. 7, pp. 815-840.
- 10) Mojirsheibani, M. (2000). A Kernel-Based Combined Classification Rule. *Statistics & Probability Letters*. 2000, Vol. 48, pp. 411-419.
- 11) Rao, R. C. (1982). *Modele liniowe statystyki matematycznej*. Warszawa: PWN.
- 12) StatSoft, Inc. (1997). *STATISTICA PL dla Windows. Tom 3*. Kraków: StatSoft Polska.