Klasyfikacja bayesowska informacji niedokładnej typu przedziałowego*

Piotr A. Kowalski†

Streszczenie: W artykule rozważane jest zagadnienie bayesowskiej klasyfikacji wielowymiarowej informacji niedokładnej typu przedziałowego, z użyciem wzorców wyznaczonych na podstawie danych określonych jednoznacznie. Do rozwiązania tak zdefiniowanego zadania zastosowana została metodyka statystycznych estymatorów jądrowych. Dodatkowo dokonywana jest eliminacja tych elementów prób wzorcowych, które mają znikomy lub wręcz negatywny wpływ na poprawność klasyfikacji. Koncepcję realizującej ten cel procedury oparto na metodzie wrażliwościowej, wzorowanej na teorii sztucznych sieci neuronowych. Dalsze polepszenie jakości klasyfikacji zostało osiągnięte przez zastosowanie algorytmu korekcji parametrów klasyfikatora.

Słowa kluczowe: analiza danych, klasyfikacja, informacja niedokładna, informacja typu przedziałowego, statystyczne estymatory jądrowe, redukcja danych, metoda wrażliwościowa w sztucznych sieciach neuronowych

1. Wstęp

Obecny gwałtowny rozwój techniki komputerowej umożliwia sukcesywne zwiększanie sprawności oraz szybkości współczesnych maszyn obliczeniowych, umożliwiając coraz częstsze wykorzystanie metod, które dotychczas były stosowane w relatywnie ograniczonym zakresie. Jedną z nich jest analiza informacji zawierającej nieokreśloność, w różnej – zależnie od uwarunkowań problemu – postaci, przykładowo niepewnej (metody statystyczne) lub nieprecyzyjnej (logika rozmyta).

Ostatnio w wielu zastosowaniach notuje się wzrost zainteresowania analizą przedziałową. Podstawą tej koncepcji jest założenie, że jedyną posiadaną informację

---

* Praca doktorska obroniona 23 listopada 2009 r. w Instytucie Badań Systemowych Polskiej Akademii Nauk. Promotor pracy: prof. dr hab. inż. Piotr Kulczycki.

† Katedra Automatyki i Technik Informacyjnych, Politechnika Krakowska, ul. Warszawska 24, 31-155 Kraków, e-mail: pkowal@pk.edu.pl; Instytut Badań Systemowych PAN, ul. Newelska 6, 01-447 Warszawa, e-mail: pakowal@ibspan.waw.pl.

o badanej wielkości stanowi fakt, iż spełnia ona zależność $\underline{x} \le x \le \overline{x}$, i w konsekwencji wielkość ta może być utożsamiana z przedziałem:

$$[\underline{x}, \overline{x}] \tag{1}$$

Analiza przedziałowa jest odrębną dziedziną matematyki, dysponującą swoim własnym aparatem formalnym, bazującym na aksjomatyce teorii zbiorów [9]. Podstawowym zastosowaniem analizy przedziałowej było zapewnienie wymaganej dokładności obliczeń numerycznych przez kontrolę błędów będących wynikiem zaokrągleń [1], jednak w wyniku ciągłego rozwoju dziedzina ta znajduje coraz szersze zastosowanie w inżynierii, ekonometrii i innych pokrewnych dyscyplinach [3]. Jej fundamentalną zaletą jest fakt, iż ze swej natury modeluje nieokreśloność badanej wielkości, stosując przy tym formuły najprostsze z możliwych. W wielu przypadkach analiza przedziałowa okazuje się całkowicie wystarczająca, a wymaga małego nakładu obliczeń (co umożliwia zastosowanie jej w bardzo złożonych zadaniach), jest łatwa w identyfikacji i interpretacji, jednocześnie posiadając formalizm oparty na dogodnym aparacie matematycznym.

Dynamicznemu rozwojowi podlegają także techniki informacyjne w zakresie analizy i eksploracji danych [7]. Wynika to nie tylko ze zwiększenia możliwości stosowanej tu metodyki, ale przede wszystkim z upowszechnienia dostępności realizujących je algorytmów, dotychczas będących domeną jedynie relatywnie wąskiej grupy specjalistów. Do podstawowych zagadnień w zakresie analizy i eksploracji danych należą zadania klasyfikacji i klasteryzacji.

Zadanie klasyfikacji polega na przypisaniu rozważanego elementu do jednej z wyróżnionych wcześniej grup. Są one najczęściej reprezentowane przez wzorce będące zbiorami elementów reprezentatywnych dla poszczególnych klas. Owa reprezentatywność sprawia, iż w wielu zagadnieniach – również tych, w których rozważana jest informacja zawierająca nieokreśloność – elementy definiujące wzorce są określone jednoznacznie (np. deterministyczne w ujęciu probabilistycznym, ostre w przypadku logiki rozmytej, czy też w nawiązaniu do notacji (1) spełniające równość $\underline{x} = \overline{x}$).

Jeśli wzorce nie mogą być ustalone arbitralnie w oparciu o specyfikę rozważanego problemu, to możliwe jest ich otrzymanie w procesie klasteryzacji, polegające na podzieleniu analizowanych danych na podzbiory elementów możliwie jak najbardziej podobnych do siebie wewnątrz każdego z nich i jednocześnie jak najbardziej odmiennych między poszczególnymi podzbiorami. Podzbiory te można następnie traktować jako naturalnie uzyskane wzorce.

Celem opisanych badań było stworzenie kompletnej procedury klasyfikacji informacji niedokładnej, danej w postaci wektora przedziałowego:

$$\begin{bmatrix} [\underline{x_1}, \overline{x_1}] \\ [\underline{x_2}, \overline{x_2}] \\ \vdots \\ [\underline{x_n}, \overline{x_n}] \end{bmatrix} \tag{2}$$

gdzie $\underline{x_k} \le \overline{x_k}$ dla $k = 1, 2, \ldots, n$, gdy wzorce poszczególnych klas są wyznaczane na podstawie zbiorów elementów określonych jednoznacznie, to znaczy przy:

$$\underline{x_k} = \overline{x_k} \quad \text{dla } k = 1, 2, \ldots, n \tag{3}$$

Koncepcję klasyfikacji oparto na ujęciu bayesowskim, zapewniającym minimum potencjalnych strat wynikłych z błędnych klasyfikacji. Do tak sformułowanego zadania została użyta metodyka statystycznych estymatorów jądrowych, co uniezależnia powyższą procedurę od arbitralnych założeń dotyczących postaci wzorców – ich identyfikacja stanowi integralną część prezentowanego algorytmu. Opracowana została także procedura redukcji liczności prób wzorcowych o te elementy, które mają znikomy lub negatywny wpływ na poprawność klasyfikacji. Jej koncepcję oparto o metodę wrażliwościową, wzorowaną na teorii sztucznych sieci neuronowych, natomiast zamysłem jest zwiększenie liczby poprawnych klasyfikacji oraz – przede wszystkim – szybkości obliczeń.

Poprawność prezentowanej tu metody została również sprawdzona dla przypadku, gdy wzorce poszczególnych klas otrzymywane są w wyniku klasteryzacji. Dodatkowo opracowano metodę zapewniającą dalsze polepszenie wyników klasyfikacji, uzyskane przez korektę wartości parametru wygładzania oraz intensywności procedury jego indywidualizacji. Poprawność oraz efektywność zastosowanych algorytmów została sprawdzona poprzez analizę numeryczną. Dokonano także porównania otrzymanych wyników z innymi istniejącymi metodami o analogicznych lub zbliżonych uwarunkowaniach.

2. Preliminaria – statystyczne estymatory jądrowe

Statystyczne estymatory jądrowe należą do metod nieparametrycznych. Umożliwiają one wyznaczenie i obrazową ilustrację charakterystyk rozkładu zmiennej losowej, bez informacji o jego przynależności do określonej klasy.

Niech (Ω, Σ, P) oznacza przestrzeń probabilistyczną. Niech także dana będzie n-wymiarowa zmienna losowa $X : \Omega \to \mathbb{R}^n$, której rozkład ma gęstość f. Jej estymator jądrowy $\hat{f} : \mathbb{R}^n \to [0, \infty)$ wyznacza się na podstawie m-elementowej próby losowej:

$$x_1, x_2, \ldots, x_m \tag{4}$$

i jest on zdefiniowany wzorem

$$\hat{f}(x) = \frac{1}{mh^n} \sum_{i=1}^{m} K\left(\frac{x - x_i}{h} \right), \tag{5}$$

przy czym dodatni współczynnik h jest nazywany parametrem wygładzania, natomiast mierzalną funkcję $K : \mathbb{R}^n \to [0, \infty)$ symetryczną względem zera, posiadającą w tym punkcie słabe maksimum lokalne i spełniającą warunek $\int_{\mathbb{R}^n} K(x)\,\mathrm{d}x = 1$, określa się mianem jądra.

Postać jądra K praktycznie nie wpływa na statystyczną jakość estymacji. W niniejszej pracy stosowane będzie uogólnione (jednowymiarowe) jądro Cauchy'ego:

$$K(x) = \frac{2}{\pi(x^2+1)^2}, \tag{6}$$

w przypadku wielowymiarowym uogólnione z wykorzystaniem koncepcji jądra produktowego:

$$K(x) = \mathscr{K}(x_1) \cdot \mathscr{K}(x_2) \cdot \ldots \cdot \mathscr{K}(x_n), \tag{7}$$

gdzie

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}, \tag{8}$$

natomiast $\mathscr{K}$ oznacza tu jednowymiarowe jądro (6). Jeżeli – jak to najczęściej ma miejsce w praktycznych zagadnieniach – dopuszcza się różne parametry wygładzania dla poszczególnych współrzędnych, oznaczane odpowiednio jako h_1, h_2, …, h_m, to definicja (5) przyjmuje postać

$$\hat{f}(x) = \frac{1}{mh_1h_2 \ldots h_n} \sum_{i=1}^{m} \mathscr{K}\left(\frac{x_1 - x_{i,1}}{h_1}\right) \mathscr{K}\left(\frac{x_2 - x_{i,2}}{h_2}\right) \ldots \mathscr{K}\left(\frac{x_n - x_{i,n}}{h_n}\right), \tag{9}$$

przy czym dodatkowo:

$$x_i = \begin{bmatrix} x_{i,1} \\ x_{i,2} \\ \vdots \\ x_{i,n} \end{bmatrix} \quad \text{dla } i = 1, 2, \ldots, m \tag{10}$$

Wartość parametru wygładzania w praktyce wyznacza się na podstawie dostępnych algorytmów, stosując posiadaną próbę losową (4). W zadaniach aplikacyjnych wprowadzane są ponadto dodatkowe procedury polepszające właściwości estymatora i dopasowujące jego cechy do badanej rzeczywistości. W niniejszych badaniach stosowana była modyfikacja parametru wygładzania, dzięki której obszary, w których estymator przyjmuje małą wartość (szczególnie tzw. „ogony") zostają dodatkowo „wygładzone", w przeciwieństwie do fragmentów jego dużych wartości (zwłaszcza w otoczeniu wartości modalnych), gdzie polepsza się charakteryzacja specyficznych cech rozkładu. Definicja (5) przyjmuje wówczas postać:

$$\hat{f}(x) = \frac{1}{mh_1h_2 \ldots h_n} \sum_{i=1}^{m} \frac{1}{s_i^n} \mathscr{K}\left(\frac{x_1 - x_{i,1}}{h_1 s_i}\right) \mathscr{K}\left(\frac{x_2 - x_{i,2}}{h_2 s_i}\right) \ldots \mathscr{K}\left(\frac{x_n - x_{i,n}}{h_n s_i}\right), \tag{11}$$

przy czym dodatnie stałe s_i stanowią parametry modyfikujące.

Szczegółowy opis metodyki statystycznych estymatorów jądrowych można znaleźć w książce [6].

3. Klasyfikacja informacji typu przedziałowego

Najpierw rozważony zostanie przypadek jednowymiarowy, to znaczy dla $n=1$.

Niech zatem dana będzie wielkość poddana procedurze klasyfikacji (2), dla rozpatrywanego teraz przypadku reprezentowana przez (jednowymiarowy) przedział:

$$[\underline{x},\overline{x}], \tag{12}$$

przy czym $\underline{x} \le \overline{x}$; jeśli $\underline{x} = \overline{x}$, to otrzymuje się klasyczny przypadek wielkości określonej jednoznacznie. Przyjmijmy ponadto, że zbiory liczb rzeczywistych:

$$x_1^1, x_2^1, \ldots, x_{m_1}^1 \tag{13}$$

$$x_1^2, x_2^2, \ldots, x_{m_2}^2 \tag{14}$$

$$\vdots$$

$$x_1^J, x_2^J, \ldots, x_{m_J}^J \tag{15}$$

reprezentują kolejno J wyróżnionych klas. Wprowadzony dodatkowo w powyższym zapisie górny indeks charakteryzuje przynależność elementu do konkretnej klasy. Jak wspomniano, zadanie klasyfikacji polega na wskazaniu, do której z nich należy przyporządkować badany element (12).

Niech teraz $\hat{f}_1$, $\hat{f}_2, \ldots, \hat{f}_J$ oznaczają estymatory jądrowe gęstości rozkładu probabilistycznego, wyznaczone kolejno w oparciu o zbiory (13)–(15) traktowane jako próby losowe – krótki opis metodyki konstruowania tego typu estymatorów przedstawiono w poprzedniej sekcji niniejszego artykułu. Zgodnie z klasycznym (tj. dotyczącym informacji określonej jednoznacznie) ujęciem bayesowskim, zapewniającym minimum potencjalnych strat wynikłych z błędnych klasyfikacji, jeżeli liczności $m_1, m_2, \ldots, m_J$ są proporcjonalne do „częstości" pojawiania się elementów z poszczególnych klas, to klasyfikowany element $\tilde{x} \in \mathbb{R}$ należy zaliczyć do tej klasy, dla której wartość:

$$m_1 \hat{f}_1(\tilde{x}),\ m_2 \hat{f}_2(\tilde{x}), \ldots, m_J \hat{f}_J(\tilde{x}) \tag{16}$$

okazuje się największa. W przypadku informacji typu przedziałowego reprezentowanej przez element (12), można przyjąć, że klasyfikowany element zalicza się do tej klasy, dla której największe jest wyrażenie:

$$\frac{m_1}{\overline{x}-\underline{x}} \int_{\underline{x}}^{\overline{x}} \hat{f}_1(x)\,\mathrm{d}x,\ \frac{m_2}{\overline{x}-\underline{x}} \int_{\underline{x}}^{\overline{x}} \hat{f}_2(x)\,\mathrm{d}x, \ldots,\ \frac{m_J}{\overline{x}-\underline{x}} \int_{\underline{x}}^{\overline{x}} \hat{f}_J(x)\,\mathrm{d}x. \tag{17}$$

Jeśli rozważyć przejście graniczne przy $\bar{x} \to \tilde{x}$ oraz $\underline{x} \to \tilde{x}$ dla ustalonego $\tilde{x} \in \mathbb{R}$, to z uwagi na ciągłość funkcji *K* danej wzorem (6), w konsekwencji implikującej ciągłość estymatora (5), otrzymuje się:

$$\lim_{\substack{\underline{x}\to\tilde{x}\\ \bar{x}\to\tilde{x}}} \frac{1}{\bar{x}-\underline{x}} \int_{\underline{x}}^{\bar{x}} \hat{f}_j(x)\,\mathrm{d}x = \hat{f}_j(\tilde{x}) \qquad \text{dla } j = 1, 2, \ldots, J. \tag{18}$$

Wyrażenia wyszczególnione w formule (18) redukują się zatem do klasycznych postaci (16).

W formule (17) dodatnia stała $1/(\bar{x}-\underline{x})$ może być pominięta jako nieistotna w problemie optymalizacyjnym, a zatem formuła ta jest równoważna postaci:

$$m_1 \int_{\underline{x}}^{\bar{x}} \hat{f}_1(x)\,\mathrm{d}x\,, \quad m_2 \int_{\underline{x}}^{\bar{x}} \hat{f}_2(x)\,\mathrm{d}x\,, \ldots\,, \quad m_J \int_{\underline{x}}^{\bar{x}} \hat{f}_J(x)\,\mathrm{d}x\,. \tag{19}$$

Co więcej, dla dowolnego $j = 1, 2, \ldots, J$ można zapisać:

$$\int_{\underline{x}}^{\bar{x}} \hat{f}(x)\,\mathrm{d}x = \hat{F}(\bar{x}) - \hat{F}(\underline{x})\,, \tag{20}$$

gdzie:

$$\hat{F}(x) = \int_{-\infty}^{x} \hat{f}(y)\,\mathrm{d}y\,. \tag{21}$$

Uwzględniając zależność (21) z podstawieniem wzorów (11) (dla $n = 1$) oraz (6), można wyliczyć, iż:

$$\hat{F}(x) = \sum_{i=1}^{m} \left[\frac{(x^2 - 2xx_i + x_i^2 + h^2 s_i^2)\; arctg\left(\dfrac{x - x_i}{s_i h}\right) + h s_i (x - x_i)}{x^2 - 2xx_i + x_i^2 + h^2 s_i^2} + \frac{\pi}{2} \right], \tag{22}$$

przy czym ponownie dodatnia stała $1/m\pi$ została pominięta. Powyższe kompletuje algorytm klasyfikacji dla przypadku jednowymiarowego.

Ostatecznie należy przyjąć, że klasyfikowany element zalicza się do tej klasy, dla której największe jest odpowiadające mu wyrażenie zawarte w formule (19), przy czym występującą tam całkę można dla każdego $j = 1, 2, \ldots, J$ efektywnie wyliczyć korzystając ze wzorów (20) i (22).

Przedstawiona koncepcja może być w naturalny sposób uogólniona na przypadek wielowymiarowy, to znaczy gdy $n > 1$. I tak, jeśli informacja typu przedziałowego jest reprezentowana przez wektor przedziałowy:

$$\begin{bmatrix} [\underline{x_1}, \overline{x_1}] \\ [\underline{x_2}, \overline{x_2}] \\ \vdots \\ [\underline{x_n}, \overline{x_n}] \end{bmatrix}, \tag{23}$$

a zbiory (13)–(15) zawierają elementy przestrzeni $\mathbb{R}^n$, to można przyjąć, że klasyfikowany element zalicza się do tej klasy dla której największe jest wyrażenie:

$$m_1 \int_E \hat{f}_1(x)\,\mathrm{d}x\,, \quad m_2 \int_E \hat{f}_2(x)\,\mathrm{d}x\,, \ldots, \quad m_J \int_E \hat{f}_J(x)\,\mathrm{d}x\,. \tag{24}$$

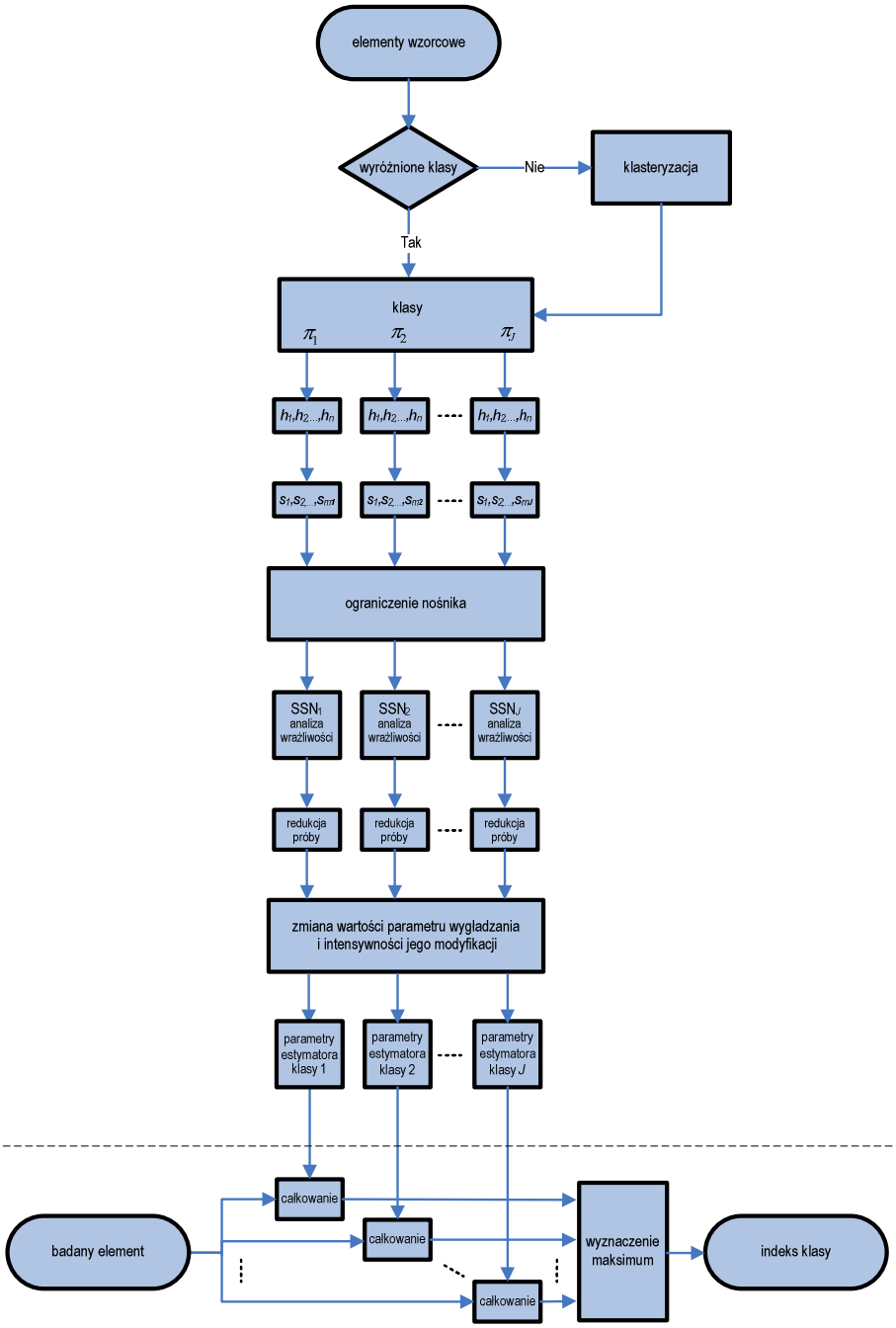
gdzie $E = [\overline{x_1}, \underline{x_1}] \times [\overline{x_2}, \underline{x_2}] \times \ldots \times [\overline{x_n}, \underline{x_n}]$. Nieco inny jest zatem algorytm obliczania występujących powyżej całek. Jednak dzięki własnościom stosowanego tu jądra produktowego, dla dowolnie ustalonego j = 1, 2, …, J, prawdziwa jest następująca zależność:

$$\int_E K(x)\,\mathrm{d}x = [\mathscr{F}(\overline{x_1}) - \mathscr{F}(\underline{x_1})][\mathscr{F}(\overline{x_2}) - \mathscr{F}(\underline{x_2})] \ldots [\mathscr{F}(\overline{x_n}) - \mathscr{F}(\underline{x_n})]\,. \tag{25}$$

gdzie $\mathscr{F}$ oznacza pierwotną postać funkcji K wprowadzonej zależnością (7). Uwzględniając definicję estymatora jądrowego z jądrem produktowym (11) oraz analityczną postać funkcji pierwotnej zawartą we wzorze (22), powyższe kompletuje procedurę klasyfikacji informacji niedokładnej typu przedziałowego, także w przypadku wielowymiarowym.

Przedstawiona procedura klasyfikacji byłaby oczywiście identyczna dla przypadku, gdy próby wzorcowe (13)–(15) są uzyskiwane za pomocą procedury klasteryzacji.

Kompletny, przedstawiony algorytm klasyfikacji zilustrowano w formie schematu blokowego (rys. 1). Wyróżnione są tam obie zasadnicze fazy opisywanej procedury: konstrukcja klasyfikatora oraz sam proces wskazania przynależności badanego elementu do wyróżnionej klasy. Pierwsza z nich, ukazana w górnej części rysunku, pokazuje poszczególne procedury począwszy od wczytania elementów wzorcowych, a skończywszy na wyznaczeniu kompletnego zbioru parametrów używanych w późniejszym procesie wskazywania konkretnej klasy. Faza ta zawiera ewentualne wyróżnienie klas przez klasteryzację, a także wyznaczenie wartości parametrów estymatorów jądrowych reprezentujących wzorce poszczególnych klas, jak również ewentualne zastosowanie procedury ograniczenia nośnika przedstawionej w [4] oraz [6]. Następnie zaznaczone są opcjonalne procedury polepszające jakość klasyfikatora, zaprezentowane w następnym punkcie niniejszego opracowania, a to: dokonywana odrębnie dla każdej klasy synteza sztucznej sieci neuronowej wraz z analizą wrażliwościową na dane uczące, przeznaczona do dokonania redukcji próby wzorcowej, jak również korekta wartości parametru wygładzania i zmiana intensywności jego modyfikacji, wykonywane w celu polepszenia jakości klasyfikacji. Ostatecznie otrzymuje się kompletny zbiór parametrów używanych w drugiej fazie – procesie wskazania konkretnej klasy, do której zaliczyć należy badany element.



Rys. 1. Schemat blokowy algorytmu klasyfikacji informacji przedziałowej
Fig. 1. The flow chart of the algorithm for interval information classification

4. Procedury zwiększające jakość klasyfikacji

W praktyce, niektóre elementy prób losowych (13)–(15), stanowiących wzorce poszczególnych klas, mogą mieć znikomy lub wręcz negatywny – z punktu widzenia poprawności procesu klasyfikacji – wpływ na jakość otrzymywanych wyników. Ich usunięcie powinno zatem skutkować redukcją liczby błędnych wskazań, a także zmniejszeniem czasu obliczeń. W tym celu zastosowana została metoda wrażliwościowa danych uczących, wzorowana na teorii sztucznych sieci neuronowych.

W celu realizacji procedury redukcji prób wzorcowych (13)–(15), konstruowane są – dla każdej z rozważanych klas – odrębne sztuczne sieci neuronowe. Konstruowana sieć jest trójwarstwowa, jednokierunkowa o m wejściach (odpowiadających poszczególnym elementom próby losowej), warstwie ukrytej o liczności równej najmniejszej liczbie naturalnej niemniejszej niż $\sqrt{m}$, a także jednym neuronie wyjściowym. Sieć była uczona z użyciem próby danych uczących, złożonej z wartości poszczególnych jąder dla kolejnych elementów próby losowej. Po zakończeniu procesu uczenia powyższa sieć została poddana procedurze analizy wrażliwościowej.

Na podstawie przedstawionych algorytmów neuronowych obliczane są współczynniki wrażliwości, określające w sposób pośredni charakter elementów wchodzących w skład próby losowej (13)–(15). W wyniku licznych badań empirycznych rekomenduje się usunięcie (redukcję) z próby wybranych elementów o charakterze redundantnym oraz typu odosobnionego, które wykazują najmniejszą wrażliwość w algorytmie neuronowym.

Więcej informacji o przytoczonym algorytmie można znaleźć w publikacji [5] oraz [8].

W literaturze przedmiotowej można znaleźć opinię, iż klasyczne uniwersalne metody wyznaczania wartości parametru wygładzania estymatora jądrowego nie są właściwymi dla zagadnienia klasyfikacji. W celu dalszego zwiększenia poprawności procesu klasyfikacji, w niniejszej pracy zaproponowano wprowadzenie $n + 1$ współczynników korekcyjnych wartości parametrów wygładzania dla poszczególnych współrzędnych oraz parametru definiującego intensywność procedury modyfikacji, wobec wartości optymalnych wyznaczonych w oparciu o kryterium minimalizacji błędu średniokwadratowego. Następnie za pomocą pełnego przeszukiwania, o stosunkowo dużej wartości dyskretyzacji siatki, znajduje się punkty najkorzystniejsze w sensie najmniejszej liczby błędnych klasyfikacji. Końcową fazą jest procedura optymalizacji statycznej w przestrzeni $(n + 1)$-wymiarowej, przy czym jako punkty początkowe przyjmuje się wyznaczone powyżej węzły siatki.

Wartość funkcji oceny jakości klasyfikacji z uwzględnieniem wartości korygujących wyznacza się z użyciem klasycznej metody *leave-one-out*. Powyższa funkcja przyjmuje wartości całkowite i dlatego nie jest możliwe stosowanie, popularnych w optymalizacji statycznej, metod gradientowych. Dlatego też użyty został algorytm Hooka-Jeevesa [2].

Warto zauważyć, iż wprowadzona w niniejszej pracy do problemu klasyfikacji procedura modyfikacji parametru wygładzania generalnie polepsza jakość klasyfikacji, a w powyższym algorytmie została dostosowana do ogólnego schematu kosztem jedynie powiększenia wymiaru przestrzeni wektorów korygujących o 1.

Także zastosowanie metody *leave-one-out* wydaje się korzystne, gdyż w przeciwieństwie do procedur opartych na dodatkowych próbach walidacyjnych, nie zmniejsza ona liczności wzorców.

Więcej informacji na temat powyższego algorytmu można znaleźć w [8].

5. Numeryczna weryfikacja algorytmu

Weryfikacja poprawności opracowanego w niniejszej pracy algorytmu klasyfikacji została potwierdzona za pomocą symulacji numerycznej. Poniżej przedstawione zostaną wnioski dla danych uzyskanych z użyciem generatora liczb pseudolosowych, o rozkładzie normalnym przy zadanym wektorze wartości oczekiwanej i macierzy kowariancji (lub ich kombinacje liniowe), otrzymywane z zaimplementowanego wielowymiarowego generatora rozkładu normalnego, opartego na koncepcji Boxa-Mullera.

Ocenę jakości przedstawionej tu metody uzyskano przez generowanie ciągów pseudolosowych o założonym rozkładzie oraz analizę poprawności wyników procedury klasyfikacji zarówno dla danych typu przedziałowego, jak i – w celach porównawczych – jednoznacznych.

Poniżej zostanie przedstawiony i skomentowany przykładowy, typowy jednowymiarowy przypadek, gdy próby reprezentujące dwa wzorce są 50-elementowe, uzyskiwane z generatorów liczb pseudolosowych o rozkładach normalnych N(0, 1) oraz N(2, 1). Warto zwrócić uwagę, iż teoretyczny punkt podziału znajduje się w odległości zaledwie jednego odchylenia standardowego od obu wartości oczekiwanych. Z kolei klasyfikowane elementy otrzymywano przez wygenerowanie pierwszej liczby pseudolosowej, a także drugiej – uzyskanej z generatora o rozkładzie jednostajnym – stanowiącej o ulokowaniu tej pierwszej wewnątrz przedziału o arbitralnie założonej długości. Powyższe reprezentuje informację typu przedziałowego w przypadku, gdy brak jest przesłanek o symetrii rozważanej niedokładności, aczkolwiek jej wielkość jest stała. Taka interpretacja wydaje się najbliższa większości praktycznych przypadków aplikacyjnych analizy przedziałowej. Klasyfikacji poddano sto zbiorów zawierających po 1000 elementów pozyskanych z każdego wzorca. Uzyskane z użyciem tej metody uśrednione wyniki zostały pokazane w tab. 1.

Tab. 1. Wyniki weryfikacji numerycznej w przypadku wzorców N(0, 1) i N(2, 1)
Tab. 1. The results of the numerical verification for patterns N(0, 1) and N(2, 1)

m \ **długość**	**0,00**	**0,10**	**0,25**
10	0,1713 ±0,0257	0,1720 ±0,0215	0,1720 ±0,0215
20	0,1655 ±0,0160	0,1669 ±0,0175	0,1669 ±0,0176
50	0,1602 ±0,0126	0,1605 ±0,0124	0,1606 ±0,0122
100	0,1596 ±0,0122	0,1601 ±0,0112	0,1602 ±0,0112
200	0,1596 ±0,0123	0,1602 ±0,0112	0,1604 ±0,0112
500	0,1591 ±0,0125	0,1595 ±0,0114	0,1596 ±0,0111
1000	0,1579 ±0,0140	0,1584 ±0,0131	0,1588 ±0,0131

0,50	**1,00**	**2,00**
0,1723 ±0,0215	0,1729 ±0,0214	0,1761 ±0,0208
0,1672 ±0,0174	0,1680 ±0,0171	0,1713 ±0,0161
0,1609 ±0,0122	0,1617 ±0,0116	0,1652 ±0,0108
0,1604 ±0,0113	0,1615 ±0,0110	0,1650 ±0,0098
0,1609 ±0,0111	0,1618 ±0,0107	0,1650 ±0,0091
0,1602 ±0,0110	0,1613 ±0,0101	0,1647 ±0,0088
0,1591 ±0,0127	0,1603 ±0,0118	0,1637 ±0,0098

Przeprowadzono również badania, gdy szerokość przedziału była ustalana losowo, analogicznie do metody stosowanej w pracy [10]. Przyjęto rozkład jednostajny na przedziale [0, 2]. Zgodnie z oczekiwaniami, otrzymane wyniki były porównywalne w relacji do przedstawionych powyżej dla ustalonych długości przedziału $[\underline{x}, \overline{x}]$.

Podobne wyniki uzyskiwano dla innych badanych rodzajów wzorców o złożonym wielomodalnym charakterze, również w liczbie większej niż dwa, także w przypadku wielowymiarowym. Zgodnie z charakterem nieparametrycznej metody estymatorów jądrowych, sytuacje takie traktowane są bowiem identycznie jak przedstawiony powyżej klasyczny przypadek jednowymiarowego rozkładu normalnego, nie zmieniając w zasadniczy sposób istoty proponowanej metodyki.

Dodatkowo w pracy [4] przedstawione zostały testy przeprowadzone dla opracowanej metody klasyfikacji informacji przedziałowej, oparte na danych „benchmarkowych" i rzeczywistych. Ze względu na specyfikę jej uwarunkowań, praktycznie nie znaleziono tego typu danych w ogólnie dostępnych repozytoriach i na stronach internetowych, a istniejących kilka przykładów nie mogło być zastosowanych w proponowanej procedurze klasyfikacji ze względu na bardzo małą liczność próby. W związku z powyższym dane przedziałowe, użyte w dalszych testach, zostały otrzymane na podstawie pozyskanych z repozytoriów danych jednoznacznych i poddane procesowi „uprzedziałowienia" w identyczny sposób, jak opisano na początku niniejszej sekcji.

We wszystkich przeprowadzonych badaniach numerycznych powiększanie liczności wzorców skutkowało zmniejszaniem zarówno średniej wartości błędu klasyfikacji, jak i jego odchylenia standardowego, co w praktyce umożliwia sukcesywne polepszanie jakości klasyfikacji w miarę pozyskiwania nowych danych. Ponadto, wraz ze zwiększającą się długością przedziału, błąd klasyfikacji wzrasta – do pewnej granicy uzasadnionej strukturą danych – w stopniu nieznacznym.

Powyższe wnioski są warte podkreślenia z aplikacyjnego punktu widzenia. Wskazują bowiem, iż możliwe jest zwiększanie jakości klasyfikacji w miarę powiększania dostępnej informacji w postaci liczniejszych wzorców oraz dokładniejszego klasyfikowanego elementu przedziałowego. W praktycznych zagadnieniach konieczne staje się zatem ustalenie kompromisu między ilością dostępnych danych a jakością otrzymywanych wyników. Poprawność prezentowanej tu metody została

również sprawdzona dla przypadku braku predefiniowanych klas, czyli gdy poszczególne wzorce otrzymywane były w wyniku klasteryzacji metodą *k-means*.

Podsumowanie

W niniejszej pracy przedstawiono kompletny algorytm bayesowskiej – a zatem zapewniającej minimum spodziewanej wartości strat – klasyfikacji wielowymiarowej informacji niedokładnej typu przedziałowego, gdy wzorce poszczególnych klas są wyznaczane na podstawie zbiorów elementów określonych jednoznacznie. Dodatkowo zaproponowano dwie opcjonalne procedury usprawniające i polepszające jakość procesu klasyfikacji: redukcję liczności próby oraz indywidualizację wartości poszczególnych parametrów. Procedurę uzupełniono o zagadnienie ograniczenia nośnika.

Z punktu widzenia złożoności obliczeniowej warto podkreślić dwuetapowość przedstawionej tu metody. Czasochłonne algorytmy konstrukcji klasyfikatora są wykonywane jednorazowo we wstępnej fazie badań. Procedura redukcji prób może odbywać się nieregularnie, w miarę wolnej mocy obliczeniowej systemu komputerowego. Sama klasyfikacja informacji niedokładnej jest dokonywana w relatywnie krótkim czasie, co w wielu zastosowaniach może mieć istotne praktyczne znaczenie. Uzyskano to w znacznym stopniu dzięki wyznaczeniu analitycznej postaci stosowanych formuł.

Przeprowadzone badania numeryczne w pełni potwierdziły pozytywne właściwości przedstawionej metody. Przeprowadzono je z wykorzystaniem danych pseudolosowych oraz rzeczywistych i „benchmarkowych". W szczególności otrzymane wyniki wskazują, iż opracowany algorytm klasyfikacji może być z powodzeniem stosowany wobec danych z klasami nieseparowalnymi o złożonych wzorcach wielomodalnych, a nawet składających się z odrębnych podzbiorów ulokowanych naprzemiennie. Uzyskano to dzięki użyciu metodyki statystycznych estymatorów jądrowych, co uniezależnia powyższą procedurę od postaci wzorców – ich identyfikacja stanowi integralną część prezentowanego algorytmu. Jak wykazała weryfikacja numeryczna, algorytm ma korzystne własności także w przypadku wielowymiarowym, bez konieczności znacznego powiększania liczności próby wzorcowej, co w pewnym stopniu dystansuje go od „przekleństwa wielowymiarowości", stanowiącej o potęgowym wzroście wymagań dotyczących liczności próby w zależności od wymiarowości przetwarzanych danych.

Zagadnienie klasyfikacji informacji przedziałowej na podstawie danych jednoznacznych można ilustracyjnie zinterpretować przykładem, gdy wzorce stanowią konkretne, precyzyjnie pomierzone dane, natomiast przedziały reprezentują ograniczenia w planach, oszacowaniach lub trudnych do przeprowadzenia pomiarach. W szczególności za próbę wzorcową można uznać bardzo dokładne pomiary, w których błędy są praktycznie pomijane, a klasyfikowany przedział reprezentuje pomiar z innego urządzenia o znacznie gorszej jakości pomiaru lub przeprowadzonego w istotnie mniej korzystnych warunkach. Odmiennym przykładem zastosowania przedmiotowej klasyfikacji jest możliwość potraktowania danych jednoznacznych jako konkretnych informacji z przeszłości, przykładowo temperatury lub kursów walut, natomiast klasyfikowany element reprezentuje prognozę o ograniczonej z natury niedokładności.

Bibliografia

1. Alefeld G., Hercberger J., *Introduction to Interval Computations*, Academic Press, New York 1986.
2. Findeisen W., Szymanowski J., Wierzbicki A., *Teoria i metody obliczeniowe optymalizacji*, PWN, Warszawa 1980.
3. Jaulin L., Kieffer M., Didrit O., Walter E., *Applied Interval Analysis*, Springer, Berlin 2001.
4. Kowalski P.A., *Klasyfikacja bayesowska informacji niedokładnej typu przedziałowego*. PhD Thesis, Instytut Badań Systemowych PAN, Warszawa 2009.
5. Kowalski P.A., Kulczycki P., *Data Sample Reduction for Classification of Interval Information using Neural Network Sensitivity Analysis*, "Lecture Notes in Artificial Intelligence", Springer-Verlag, 2010, vol. 6304, 271–272.
6. Kulczycki P., *Estymatory jądrowe w analizie systemowej*, WNT, Warszawa 2005.
7. Kulczycki P., Hryniewicz O., Kacprzyk J. (red.), *Techniki informacyjne w badaniach systemowych*, WNT, Warszawa 2007.
8. Kulczycki P., Kowalski P.A., *Bayes classification of imprecise information of interval type*, "Control and Cybernetics", 2011, vol. 40, no. 1, 101–123.
9. Moore R.E., *Interval Analysis*, Prentice-Hall, Englewood Cliffs 1966.
10. Souza De R.M.C.R., Carvalho De F.A.T., *Dynamic clustering of interval data based on adaptive Chebyshev distances.*, "Electronics Letters", 2004, vol. 40, 658–660.

Bayes classification of imprecise information of interval type

Abstract: The subject of the investigations presented here was a classification procedure, where the classified element is given as the interval vector, but the data representing each class consists of elements defined uniformly. The concept of classification was based on the Bayes approach and statistical kernel estimators methodology. The procedure has also been worked out for reducing samples, based on the sensitivity method inspired by neural networks. The sensitivity analysis allows the removal of neutral or even harmful (from the point of view of proper results of the classification procedure) elements of the pattern set. Better quality of classification was achieved with an algorithm for the correction of classifier parameter values.

Keywords: data analysis, classification, imprecise information, interval type information, statistical kernel estimators, reduction in pattern size, sensitivity method for artificial neural networks