Symulacja w Badaniach i Rozwoju
Vol. 3, No. 2/2012

Zenon A. SOSNOWSKI
Politechnika Białostocka, Wydział Informatyki
ul. Wiejska 45a , 15-351 Białystok
E-mail: z.sosnowski@pb.edu.pl

Badanie klasyfikatora rozmytego z wykorzystaniem entropii rozmytej

1 Wstęp

W dzisiejszym świecie istnieje nieustanna potrzeba rozwijania jak najbardziej efektywnych metod umożliwiających, a także ułatwiających operowanie danymi. Jednym z najistotniejszych problemów z tym związanych jest klasyfikacja danych. Metody klasyfikacyjne znajdują zastosowanie niemal w każdej dziedzinie życia, np.: w medycynie, biznesie, itp. Szczególnie prężnie rozwija się w ostatnich latach dziedzina klasyfikacji rozmytej. Wyróżniającą cechą klasyfikatorów rozmytych jest to, że opierają się one na logice rozmytej, która dopuszcza istnienie wartości pośrednich pomiędzy prawdą a fałszem, przez co jej możliwości są większe niż tradycyjnej logiki dwuwartościowej. Klasyfikację i klasyfikatory rozmyte stosuje się zwłaszcza w sytuacjach, w których nie mamy wystarczającej wiedzy o modelu matematycznym lub też musimy podjąć decyzję w okolicznościach nie do końca nam znanych. Aktualny stan badań teoretycznych i doświadczalnych wskazuje na potrzebę prowadzenia dalszych prac w kierunku doskonalenia modelowania tego typu klasyfikatorów, jak również weryfikacji eksperymentalnych. Celem pracy jest zbadanie i porównanie metod klasyfikacji rozmytej opartych o entropię rozmytą. Badania przeprowadzone zostały w oparciu o architekturę klasyfikatora opisanego w [1].

2 Opis klasyfikatora rozmytego opartego na entropii rozmytej

W tworzeniu klasyfikatora dwie kwestie są istotne – zmniejszanie czasu klasyfikacji i zwiększanie jej jakości. Do konstruowania klasyfikatorów rozmytych używano różnych metod. Jedną z prostszych było stosowanie do tego celu sieci neuronowych. Trenowanie takiej sieci jednakże, zajmuje dużo czasu. Wyniki są trudne w analizie, a w konsekwencji trudno jest poprawić jakość wytrenowanej sieci. By rozwiązać te problemy i stworzyć odpowiednie klasyfikatory, zaproponowanych zostało wiele metod opartych na analizie obszarów rozmytych. W ogólności, "kształty" takich obszarów rozmytych mogą być zaklasyfikowane do następujących typów:

1. obszary hipersześcienne, których granice równoległe są do osi wejściowych,
2. obszary elipsoidalne,
3. obszary wielościenne, których granice wyrażone są przez liniową kombinację zmiennych wejściowych.

Klasyfikator używający obszarów hipersześciennych jest łatwiejszy i szybszy od tych, które używają obszarów elipsoidalnych, czy też wielościennych.

Badany klasyfikator rozmyty oparty jest na nieprzecinających się obszarach hipersześciennych. Zamiast reguł rozmytych używana jest tu miara entropii rozmytej

[1]. Faktyczny rozkład wzorców klasyfikacyjnych odzwierciedlany jest przez zliczenie entropii rozmytej poszczególnych wymiarów. Obszar decyzyjny może być automatycznie poprawiony w oparciu o miarę entropii rozmytej. Ze względu na zdolność klasyfikatora do wykrywania faktycznego rozkładu wzorców, wygenerowane regiony decyzyjne mogą zostać w efektywny sposób zmniejszone.

Najistotniejszą procedurą dla systemu klasyfikacji jest podział przestrzeni wzorców na obszary decyzyjne. Po ustaleniu obszarów decyzyjnych możemy ich używać do klasyfikacji nieznanych wzorców. Podział na obszary decyzyjne jest częścią procesu uczeniu klasyfikatora, gdyż o postaci tych obszarów decydują wzorce uczące.

W badanym, opartym na entropii rozmytej klasyfikatorze rozmytym, regiony decyzyjne domykane są powierzchniami utworzonymi z każdego z wymiarów. Powierzchnie są określone przez rozkład wzorców wejściowych. Powierzchnie każdej z tych podprzestrzeni są odpowiednio równoległe do każdego z wymiarów, tzn. powierzchnie są rozszerzane z granic funkcji przynależności (nazwanych interwałów) na każdym z wymiarów.

Aby utworzyć powierzchnie na każdym z wymiarów, lub też - co jest równoważne – by wygenerować kilka trójkątnych funkcji przynależności dla każdego z atrybutów rzeczywistych (proces nazywany również dyskretyzacją atrybutów [2,3, 4,5,6,7]) należy rozważyć kilka problemów. Najpierw trzeba określić liczbę przedziałów na każdym z wymiarów. Następnie musi zostać wyliczony środek i szerokość każdego z interwałów. W celu określenia właściwej liczby przedziałów używamy entropii rozmytej, natomiast środki interwałów znajdujemy używając algorytmu grupowania K-średnich. Gdy już zostaną ustalone środki przedziałów, to łatwo jest wyznaczyć ich szerokości.

Z powyższego opisu możemy podsumować badany klasyfikator w następujących czterech krokach:

1. ustalenie liczby przedziałów na każdym z wymiarów,
2. ustalenie pozycji przedziałów, tzn. ustalenie środka i szerokości każdego przedziału,
3. przypisanie dla każdego z przedziałów funkcji przynależności,
4. przypisanie dla każdego obszaru decyzyjnego etykiety klasy.

Ustalenie liczby przedziałów na każdym z wymiarów

W [8,9,10] pokazano, że liczba przedziałów ma głęboki wpływ na efektywność uczenia się i dokładność klasyfikacji. W przypadku, gdy ta liczba jest zbyt duża, tj. podział jest zbyt "dokładny", proces uczenia i klasyfikacji będzie trwał zbyt długo, przy czym może wystąpić zjawisko zbyt dobrego dopasowania do danych. Z drugiej strony, jeśli liczba przedziałów jest zbyt mała, to rozmiar każdego z obszarów decyzyjnych może być zbyt duży by pasować do rozkładu wzorców wejściowych, co może osłabić wydajność klasyfikacji.

Wybór optymalnej liczby przedziałów jest jednak w literaturze rzadko poruszany. W większości przypadków jest wyznaczany w sposób dowolny lub też heurystycznie. W tym podrozdziale będziemy badać systematyczną metodę doboru właściwej liczby przedziałów. Zaproponowane kryterium oparte jest na mierze entropii

rozmytej (opisanej w [1]), ponieważ ma ona zdolność odzwierciedlania faktycznego rozkładu przestrzeni wzorców. Proces wyboru najlepszej liczby przedziałów dla każdego z wymiarów przebiega w następujących krokach:

Krok 1) Ustaw początkową liczbę przedziałów na I = 2.

Krok 2) Znajdź środki przedziałów.

W celu znalezienia środka każdego z przedziałów, a co za tym idzie jego końców i szerokości używamy algorytmu grupującego.

Krok 3) Dla każdego przedziału przypisz funkcję przynależności.

Przypisanie funkcji przynależności dla każdego z przedziałów umożliwi zastosowanie entropii rozmytej do pomiaru informacji o rozkładzie wzorców w danym przedziale.

Krok 4) Policz całkowitą entropię rozmytą wszystkich przedziałów dla I oraz I-1 przedziałów.

Obliczamy entropię rozmytą wszystkich przedziałów na każdym z wymiarów aby otrzymać informację o rozkładzie wzorców zrzutowanych na ten wymiar. Funkcja entropii rozmytej opisana została w [1].

Krok 5) Czy całkowita entropia rozmyta zmniejsza się?

Jeżeli całkowita entropia rozmyta I przedziałów jest mniejsza od tej dla

I-1 przedziałów (tzn. podział tego wymiaru na I przedziałów będzie bardziej uporządkowany niż ten dla I-1 przedziałów), wówczas dzielimy jeszcze raz

(I := I + 1) i przechodzimy do Kroku 2; w przeciwnym wypadku przejdź do Kroku 6.

Krok 6) I-1 jest optymalną liczbą przedziałów na danym wymiarze.

Skoro entropia rozmyta nie zmniejsza się, rezygnujemy z dalszego podziału tego wymiaru, a I-1 jest optymalną liczbą przedziałów na tym wymiarze.

Wyznaczanie umiejscowień przedziałów

Proces wyznaczenia umiejscowień przedziałów rozpoczyna się znajdywaniem ich punktów centralnych. Kiedy już środek przedziału jest wyznaczony, łatwo jest zdecydować o jego szerokości i granicach. Metoda wyznaczania szerokości i granic przedziału opisana jest w następnym podrozdziale. W celu znalezienia środków, używamy algorytmu K-średnich. Jest to użyteczna, nie wymagająca nadzoru metoda uczenia się oparta na Euklidesowej mierze odległości.

Załóżmy, że mamy N M – wymiarowych wektorów $V_i = (v_{i1}, v_{i2}, ..., v_{im})^T$, $i = 1, 2, ..., N$, odnoszących się do N elementów. W celu podzielenia elementów na kilka przedziałów na wymiarze j, wyciągamy najpierw N wartości z elementów zrzutowanych na ten wymiar $x_i^{(j)} = v_{ij}$, $i = 1, 2, ..., N$. Następnie używany jest algorytm grupujący K – średnich, aby dokonać zgrupowania na $x_i^{(j)}$, $i = 1, 2, ..., N$. Algorytm składa się z następujących kroków.

Krok 1) Ustawiamy początkową liczbę grup, I.

To jest procedura wyznaczania liczby I przedziałów, opisana w poprzednim podrozdziale.

Krok 2) Ustawiamy początkowe wartości środków grup.

Początkowe środki grup c_1 , c_2 , ..., c_I mogą zostać wybrane w sposób losowy spośród wartości $x_i^{(j)}$, $i = 1, 2, ..., N$. W proponowanym systemie, środek c_q dowolnie wybranej grupy q przypisywany jest następująco:

$$c_q = \frac{q-1}{I-1} \ , \ q = 1, 2, ..., I$$

Krok 3) Przypisanie etykiety grupy do każdego elementu.

Po wyznaczeniu środków grup, każdemu z elementów przypisujemy etykietę grupy, od środka której dzieli go najmniejszy dystans. Jest to środek najmniejszej odległości euklidesowej od elementu. Zatem najbliższy środek spełnia następującą miarę odległości:

$$\left|x_i^{(j)} - c_q^*\right| = \min_{1 \le q \le I} \left|x_i^{(j)} - c_q\right|$$

gdzie c_q^* jest najbliższym środkiem elementu $x_i^{(j)}$, tzn. ma najmniejszą odległość euklidesową do $x_i^{(j)}$ spośród wszystkich środków : c_1 , c_2 , ..., c_I .

Krok 4) Ponowne przeliczenie środków grup.

Ponieważ początkowe środki grup zostały wybrane w sposób losowy, musimy ponownie przeliczyć każdy ze środków poprzez wyliczenie następującego oszacowania:

$$c_q = \frac{\sum_{i=1}^{N_q} x_i^{(j)}}{N_q}$$

gdzie N_q jest całkowitą liczbą wzorców wewnątrz tej samej grupy q.

Krok 5) Czy któryś ze środków zmienił się?

Jeżeli każdy środek grupy został właściwie wyznaczony, jego ponowne przeliczenie w Kroku 4 da taki sam wynik. Jeżeli zachodzi właśnie taka sytuacja, to przerywamy proces wyznaczania środków grup, w przeciwnym wypadku przechodzimy do kroku 3.

Przypisanie funkcji przynależności

Przypisanie funkcji przynależności jest procedurą wyznaczającą funkcję przynależności każdemu z przedziałów. By móc zastosować entropię rozmytą do analizy informacji o rozkładzie wzorców w danym przedziale, należy przypisać każdemu przedziałowi odpowiednią funkcję przynależności, by pokazać stopnie przynależności elementów. Wartość przynależności elementu znajdującego się wewnątrz przedziału może być rozpatrywana jako stopień przynależności tego elementu do danego przedziału. Intuicyjnie, środek przedziału ma najwyższy stopień przynależności, natomiast wraz ze wzrostem odległości elementu od środka przedziału, jego stopień przynależności do tego przedziału maleje. Zatem najwyższą wartość przynależności "1.0" przypisujemy dla środka przedziału, natomiast najmniejszą wartość "0.0" - środkom przedziałów sąsiadujących. W naszym przypadku, aby wprowadzić ten system, używamy zbiorów rozmytych trójkątnych.

Przypisując do przedziału funkcję przynależności, musimy rozważyć trzy podstawowe przypadki:

Przypadek I) Przedział najbardziej wysunięty na lewo.

W tym przypadku, środek pierwszego przedziału c_1 na tym wymiarze jest ograniczony tylko jednym środkiem przedziału c_2 . Najwyższa wartość przynależności "1.0" tego przedziału znajduje się w punkcie c_1 , najmniejsza zaś, "0.0", znajduje się w punkcie c_2 . Dla $x = x_{min}$ wartość funkcji przynależności ustawiona jest na 0.5.

Przypadek II) Przedziały wewnętrzne.

W tym przypadku, środek wewnętrznego przedziału *ci* ograniczony jest przez środek c_{i-1} interwału po lewej stronie oraz środek c_{i+1} interwału po prawej stronie. Najwyższa wartość przynależności znajduje się w punkcie c_i , zaś najniższa – w punktach c_{i-1} oraz c_{i+1} .

Przypadek III) Przedział najbardziej wysunięty na prawo.

W tym przypadku środek ostatniego przedziału c_I na tym wymiarze jest ograniczony tylko jednym środkiem przedziału c_{I-1} . Najwyższa wartość przynależności "1.0" tego przedziału znajduje się w punkcie c_I , najmniejsza zaś, "0.0", znajduje się w punkcie c_{I-1} . Dla $x = x_{max}$ wartość funkcji przynależności ustawiona jest na 0.5.

Przypisanie etykiet klas

W celu przypisania etykiet klas dla każdego obszaru decyzyjnego możemy użyć dwóch metod: metody większościowej oraz metody szacowania entropii rozmytej.

Metoda większościowa polega na przypisaniu obszarowi decyzyjnemu tej klasy, której obiektów jest w tym obszarze najwięcej. W przypadku, gdy jest kilka takich klas to wybór dokonywany jest losowo z puli klas równolicznych w tym obszarze.

Druga metoda podobna jest do metody szacowania wartości entropii rozmytej opisanej wyżej. W odróżnieniu od poprzedniego przykładu pomiaru wartości entropii rozmytej dla przedziału, aby wyznaczyć klasę każdego z obszarów decyzyjnych liczone są wartości entropii rozmytej wzorców należących do poszczególnych klas. W istocie rzeczy, entropia rozmyta obszaru decyzyjnego może być otrzymana poprzez zsumowanie wartości entropii rozmytej poszczególnych przedziałów w każdym z wymiarów. Obszarowi decyzyjnemu przypisujemy tę klasę, która ma najniższą wartość entropii rozmytej w tym obszarze.

Przypisanie każdemu obszarowi decyzyjnemu odpowiedniej etykiety klasy kończy proces uczenia klasyfikatora.

3 Eksperymenty

Przedstawione zostaną przebieg i wyniki (Tabela 1) testów jakości badanego klasyfikatora wykonane przy użyciu aplikacji wykonanej w ramach pracy dyplomowej [11]. Badania przeprowadzone zostały na zbiorach: *Iris Database* (baza danych Irysów) oraz *Wisconsin Breast Cancer Diagnostic Database* (baza danych Wisconsin Breast

Cancer). Są to powszechnie znane zbiory danych [12] , często używane do testów jakości klasyfikatorów.

Zbiór *Iris Database* zawiera 150 elementów, z których każdy należy do jednej z trzech klas (50 elementów na każdą z klas). Baza ta posiada cztery ciągłe cechy. Zbiór *Wisconsin Breast Cancer Diagnostic Database* zawiera 699 wzorców, z których każdy należy do jednej z dwóch klas. Każdy wzorzec opisany jest za pomocą dziewięciu cech ciągłych.

Jako kryterium końcowe obrana została wartość jakości klasyfikacji klasyfikatora liczona według metody kroswalidacji Monte-Carlo z następującymi parametrami:

- odsetek obiektów losowanych w każdym kroku do zbioru treningowego: 50% ,
- odsetek obiektów losowanych w każdym kroku do zbioru testowego: 50% ,
- liczba iteracji: 50.

Dla każdego ze zbiorów rozpatrywano dwa sposoby wyznaczania klasy regionu decyzyjnego: w oparciu o tradycyjną metodę większościową, a także w oparciu o miarę entropii rozmytej.

Zbiór danych	**Jakość klasyfikacji (metoda większościowa)**	**Jakość klasyfikacji (entropia rozmyta)**
Iris	82,3466 %	38,8800 %
Wisconsin Breast Cancer Diagnostic	73.11695 %	90.9532 %

Tab. 1. Porównanie wyników klasyfikacji

Tab. 1. Comparison of the results of the classification

Analizując uzyskane wyniki można wnioskować, że jakość klasyfikatora jest warta uwagi.

Przyjrzyjmy się bliżej rezultatom pomiaru jakości klasyfikacji metodą kroswalidacji Monte-Carlo dla zbiorów danych Iris Database oraz Wisconsin Breast Cancer Diagnostic Database. Jak widać, wyniki są zróżnicowane, zarówno biorąc pod uwagę zbiór użyty do procesu testowego, jak też metodę przyporządkowywania klas regionom decyzyjnym. I tak np., dla zbioru Iris Database wynik pomiaru jakości klasyfikacji uzyskany przy metodzie większościowej jest zdecydowanie większy niż ten odpowiadający metodzie opartej na wartości entropii rozmytej. W przypadku zbioru Wisconsin Breast Cancer Diagnostic Database przewagę w stosunku do metody większościowej (73.11695 %) uzyskała metoda oparta na mierze entropii rozmytej

(90.9532 %). Nasuwa się wniosek, iż w przypadku zbioru Iris Database metoda oparta na entropii rozmytej jest w pewien sposób „czuła" na cechy redundantne, których miara entropii rozmytej jest wysoka.

4 Wnioski końcowe

Podsumowując, badany klasyfikator, zarówno w przypadku, gdy przydziela klasy regionom decyzyjnym w oparciu o miarę entropii rozmytej, jak też przy użyciu metody większościowej, daje w ogólności dość dobre rezultaty. Generowane przez niego nieprzecinające się, hipersześcienne regiony decyzyjne sprawiają, że czas klasyfikacji wzorców jest stosunkowo krótki. Mimo tego, że w niektórych przypadkach klasyfikator nie daje wysokich rezultatów, to jednak większość z przeprowadzonych badań ukazała go jako warty uwagi. Zatem możemy uznać go, wraz z zawartymi w nim mechanizmami pomiaru entropii rozmytej wzorców klasyfikacyjnych, za interesującą propozycję w przestrzeni klasyfikatorów rozmytych.

Literatura

1. Hahn-Ming Lee, Chih-Ming Chen, Jyh-Ming Chen, and Yu-Lu Jou: *An Efficient Fuzzy Classifier with Feature Selection Based on Fuzzy Entropy*, IEEE Transactions on Systems, Man, and Cybernetics—Part B: Cybernetics, 3 June 2001
2. Ching J. Y. et al.: *Class-dependent discretization for inductive learning form continuous and mixed-mode data*, IEEE Trans. Pattern Anal. MachineIntell., Lipiec 1995
3. Liu H. and Setiono R.: *Feature selection via discretization*, IEEE Trans. Knowl. Data Eng., Lipiec/Sierpień 1997
4. Jun B.H. et al.: *A new criterion in selection and discretization of attributes for the generation of decision trees*, IEEE Trans. Pattern Anal. Machine Intell., Grudzień 1997
5. Pedrycz W., Sosnowski Z.A.: *The designing of decision trees in the framework of granular data and their application to software quality models*, Fuzzy Sets & Systems, vol. 124, 2001
6. Shen H., Yang J., Wang S., and Liu X.: *Attribute weighted mercer kernel based fuzzy clustering algorithm for general non-spherical datasets*, Soft Computing, vol. 10, no. 11, 2006
7. Hsin-Chien Huang, Yung-Yu Chuang and Chu-Song Chen: *Multiple Kernel Fuzzy Clustering,* IEEE Trans. Fuzzy Syst., Volume: 20 , Issue: 1, 2012
8. Ching J.Y. et al.: *Class-dependent discretization for inductive learning form continuous and mixed-mode data*, IEEE Trans. Pattern Anal. MachineIntell., lipiec 1995
9. Nozaki K. et al.: *Adaptive fuzzy rule-based classification systems*, IEEE Trans. Fuzzy Syst., czerwiec 1996
10. Pedrycz W., Sosnowski Z.A.: *C-Fuzzy Decision Trees*, IEEE Transactions on Systems, Man and Cybernetics, Part C, Vol. 35, No 4, 2005
11. Gudwański T.: *Badanie metod klasyfikacji rozmytej opartych o selekcję cech*, praca magisterska, Politechnika Białostocka, Wydz. Informatyki, 2007
12. Merz C.J, Murphy P.M.: *UCI Repository for Machine Learning Data-Bases* [http://www.ics.uci.edu/~mlearn/MLRepository.html], Irvine, CA: University of California, Department of Information and Computer Science, 1996

Streszczenie

W pracy przedstawiono architekturę klasyfikatora rozmytego opartego na entropii rozmytej oraz zbadano jego wydajność na standardowych zestawach danych: *Iris* i *Wisconsin breast cancer*. Wyniki symulacji pokazują, że przedstawiony klasyfikator daje zadawalające wskaźniki klasyfikacji.

Słowa kluczowe: klasyfikator rozmyty, rozmyta entropia

Study of fuzzy classifier based on fuzzy entropy

Summary

In this paper, we present the architecture of fuzzy classifier based on fuzzy entropy and examine its performance on Iris and Wisconsin breast cancer data sets. Simulation results show that the presented classifier has a satisfactory classification rate.

Keywords: Fuzzy classifier; Fuzzy entropy

Praca finansowana w ramach badań statutowych Wydziału Informatyki Politechniki Białostockiej nr S/WI/2/08.