Symulacja w Badaniach i Rozwoju
Vol. 3, No. 4/2012

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

Reguły decyzyjne z rozmytą granulacją wiedzy

1 Wstęp

Wraz ze wzrostem mocy obliczeniowej komputerów w ostatnich latach można zaobserwować również znaczny wzrost zgromadzonych i udostępnianych danych. Z tego powodu potrzebne są coraz bardziej zaawansowane i doskonałe algorytmy ułatwiające ich analizę, w tym ewentualną klasyfikację [1]. Klasyfikacja jest wykorzystywana między innymi w takich dziedzinach, jak: rozpoznawanie trendów na rynkach finansowych, wspomaganie decyzji przyznawania kredytów bankowych, automatyczne rozpoznawanie obiektów w dużych bazach danych obrazów oraz w systemach medycznych w klasyfikacji schorzeń.

Jedną z bardzo popularnych pośród wszystkich dostępnych metod wykorzystywanych do zadań klasyfikacji jest indukcja drzew decyzyjnych [2]. Istotną zaletą drzew decyzyjnych jest fakt, iż skonstruowany model jest zrozumiały dla człowieka. Kolejne cechy wymienionego modelu to możliwość zastosowania zbiorów danych wielowymiarowych oraz wysoka skalowalność dla dużych zbiorów danych. Pośród wad drzew decyzyjnych można wymienić dwie podstawowe cechy. Przy analizie złożonych hipotez otrzymane drzewa mogą być bardzo złożone, ponieważ stosowane testy w znaczącej większości przypadków weryfikują wartości jedynie pojedynczych atrybutów. W rezultacie powoduje to utratę możliwości wykorzystania zależności pomiędzy zbiorami cech. Ominięcie drugiej wymienionej wady mogłoby prowadzić do konstrukcji mniej złożonego modelu klasyfikacji [3,4]. Dodatkowo, w przypadku występowania algorytmów o wartościach ciągłych, dyskretyzacja jest koniecznością, a tym samym może mieć istotny wpływ na wydajność powstałego drzewa. Algorytm badany w niniejszej pracy – klastrowo-kontekstowe rozmyte drzewa decyzyjne – w konstrukcji węzłów drzewa wykorzystuje zależności pomiędzy kompletnym zbiorem cech i jest to jego istotną zaletą w porównaniu do standardowych algorytmów drzew decyzyjnych. Dodatkowo, z powodzeniem może być stosowany w klasyfikacji zbiorów danych, gdzie zmienna decyzyjna zawiera wartości zarówno ciągłe, jak i dyskretne [3].

Celem pracy jest przedstawienie możliwie najbardziej efektywnego połączenia dwóch metod hierarchicznego grupowania rozmytego: klastrowych rozmytych drzew decyzyjnych [3] oraz kontekstowych drzew decyzyjnych [5], jak również zbadanie korzyści wynikłych z nowej metody po przeprowadzeniu szczegółowych badań i podsumowaniu otrzymanych wyników. Rezultatem badań jest również potwierdzenie, iż połączenie dwóch wymienionych wcześniej algorytmów umożliwi uzyskanie lepszych wyników niż uzyskane przez istniejącą wcześniej metodę: klastrowych rozmytych drzew decyzyjnych [3].

2 Opis metody klastrowo-kontekstowych rozmytych drzew decyzyjnych

Zadaniem metody klastrowo-kontekstowych rozmytych drzew decyzyjnych jest efektywna integracja dwóch algorytmów: klastrowych rozmytych drzew decyzyjnych [3] oraz kontekstowych drzew decyzyjnych [5]. Celem badań jest potwierdzenie lub zaprzeczenie tezy, iż połączenie tych dwóch algorytmów może dać lepsze jakościowo wyniki. W dalszej części zostaną przedstawione kolejne fazy budowy klastrowo-kontekstowego rozmytego drzewa decyzyjnego [6] oraz sposób zastosowania tej metody w zadaniu klasyfikacji.

Podział zmiennej decyzyjnej na konteksty

Podstawowym elementem klastrowo-kontekstowych rozmytych drzew decyzyjnych są konteksty, na które zostaje podzielona zmienna decyzyjna. Kolejnym krokiem jest przyporządkowanie wartości atrybutu decyzyjnego dla każdego kontekstu, mającą wartość funkcji przynależności, zawierającą się w przedziale od 0 do 1. Aby dobrać wartości parametrów funkcji przynależności, należy najpierw znaleźć wartości minimalne i maksymalne zmiennej decyzyjnej.

Ostatecznie wartości zmiennej decyzyjnej zostają podzielone na równe przedziały według liczby kontekstów. Do analizy wykorzystano trzy popularne rodzaje funkcji przynależności: trójkątną, gaussowską i trapezoidalną, które mogą być z powodzeniem stosowane w teorii zbiorów rozmytych [5]. Alternatywą dla tego wyboru mogłoby być osadzenie kontekstów w taki sposób, aby każdy z nich obejmował taką samą liczbę obiektów. Nie jest to jednak preferowane rozwiązanie, ponieważ taki podział mógłby być zdominowany przez obszary, w których jest najwięcej obiektów, co może mieć negatywny wpływ na uzyskane wyniki. Listę kroków wykonywanych w początkowej fazie algorytmu, zajmującej się inicjalizacją kontekstów, przedstawiono poniżej:

1. Znajdź wartości min i max zmiennej decyzyjnej y.
2. Oblicz wartość odległości względnej pomiędzy maksymalną wartością funkcji przynależności w danych kontekstach.
3. Dla każdego kontekstu k wyznacz parametry funkcji przynależności zależnie od wybranej funkcji zgodnie z równomiernym podziałem całego obszaru wartości zmiennej decyzyjnej pomiędzy konteksty.

Budowa drzewa klastrowo-kontekstowego

Każde z otrzymanych drzew kontekstowych charakteryzowane jest przez następujący zestaw wartości:

- parametry funkcji przynależności skojarzone z wybranym kontekstem,
- wybrana funkcja przynależności,
- liczba klastrów c,
- numer kontekstu k, którego dotyczy drzewo,
- zbiór wartości kontekstu dla obiektów przynależących do danego drzewa oznaczony przez Uc,
- zbiór wartości x obiektów, dla których funkcja przynależności jest większa od epsilon (0.001 dobrane arbitralnie) oznaczona jako Xc,

- zbiór wartości zmiennych wyjściowych obiektów przynależących do aktualnego drzewa kontekstowego.

Każde z drzew kontekstowych jest budowane z zachowaniem zasad przedstawionych w metodzie klastrowych rozmytych drzew decyzyjnych, ale dodatkowo wprowadzona jest tu wartość związana z kontekstem, określona z zastosowaniem wybranej funkcji przynależności. Należy zaznaczyć, iż do budowy drzewa kontekstowego wykorzystywany jest podzbiór obiektów z całego zbioru danych. Oznacza to, że w aktualnie budowanym kontekście brane są pod uwagę tylko obiekty, dla których wartość kontekstu powiązana ze zmienną decyzyjną jest większa od określonej wartości (wartość wybrana arbitralnie to 0.00001). Następnie macierz partycji wykorzystywana do rozbudowy każdego klastra inicjalizowana jest z użyciem wartości f_k, zgodnie ze wzorem (1), i spełnia warunek ze wzoru (2).

$$\sum_{i=1}^{c} \mathrm{u}_{ik}\,(x) = f_k, \;\; k = 1,2,\dots,N\,, \tag{1}$$

$$\boldsymbol{u_{ik}(x) = \frac{f_k}{\sum_{j=1}^{c}\left(\frac{\|x_k - v_i\|}{\|x_k - v_j\|}\right)^{\frac{2}{m-1}}}\,,\;\; m = 2\,.} \tag{2}$$

Zastosowana funkcja odległości to ważona odległość euklidesowa zdefiniowana według wzoru:

$$\|\mathrm{a} - \mathrm{b}\| = \sum_{i=1}^{n} \frac{(\mathrm{a}_i - b_i)^2}{{\sigma_i}^2}\,, \tag{3}$$

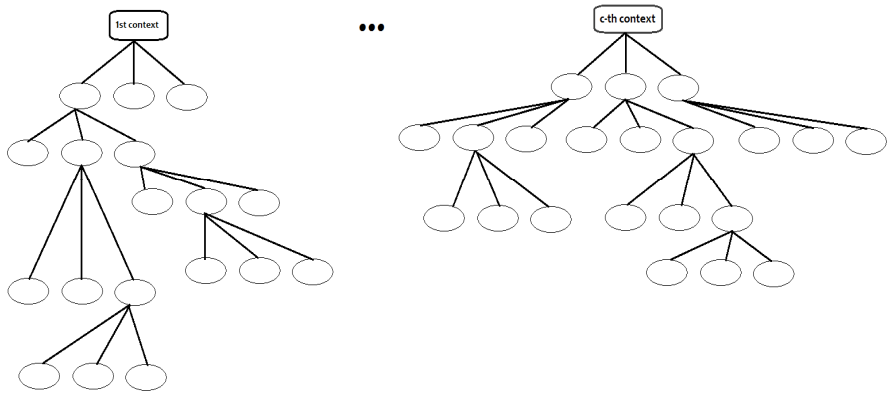
gdzie *n* oznacza liczbę atrybutów obiektu *x*.

Ogólne kroki powstałego algorytmu:

1. Mając zdefiniowaną liczbę klastrów oraz funkcję przynależności, zdefiniuj konteksty na zmiennej decyzyjnej (y).
2. Dla każdego kontekstu utwórz oddzielne rozmyte drzewo decyzyjne, którego węzeł główny zawiera zbiór obiektów, dla jakich wartość funkcji przynależności w danym kontekście jest większa niż e (0.001).
    a. Do budowy drzewa zastosuj algorytm klastrowych rozmytych drzew decyzyjnych z użyciem zmiennej warunkowej (kontekstu), jako początkowe wartości funkcji przynależności (Uc).
        i. Grupuj dane Xc (zbiór dla i-tego kontekstu).
        ii. Powtarzaj
            1. Przyporządkuj elementy Xc do odpowiednich klastrów.
            2. Wybierz węzeł z największą wartością współczynnika zróżnicowania i grupuj dane w wybranym węźle do momentu, aż zostanie spełnione kryterium stopu (współczynnik zróżnicowania bliski 0).

Klasyfikacja metodą klastrowo-kontekstowych rozmytych drzew decyzyjnych

Schemat drzew powstałych po zastosowaniu metody klastrowo-kontekstowych rozmytych drzew decyzyjnych przedstawiono na rys 1.



Rys. 1. Schemat klastrowo-kontekstowego drzewa decyzyjnego z liczbą kontekstów c i liczbą klastrów równą 3

Fig. 1. C-context decision tree with c contexts and the number of clusters equal to 3

Drzew jest dokładnie tyle, ile zostało zdefiniowanych kontekstów, dlatego istotne było określenie nowego sposobu, w jaki otrzymana struktura może być wykorzystana w trybie klasyfikacji.

Na początku obiekt jest klasyfikowany kolejno z użyciem wszystkich otrzymanych drzew kontekstowych. W każdym drzewie kontekstowym klasyfikacja odbywa się analogicznie do metody klastrowych rozmytych drzew decyzyjnych [2]. Stopniowy wybór klastra podrzędnego odbywa się z wykorzystaniem wzoru (4). Obiekt jest przypisany do klasy, jeśli wartość $u_i(\mathrm{x})$ jest większa od wartości funkcji przynależności we wszystkich pozostałych klastrach i tak samo realizowany jest wybór klastra przy wyborze ścieżki do liścia. Wartość u_i jest wyliczana zgodnie ze wzorem (4):

$$u_i(\mathrm{x}) = \frac{1}{\sum_{j=1}^{c}\left(\frac{\|\mathrm{x}-\mathrm{v}_i\|}{\|\mathrm{x}-\mathrm{v}_j\|}\right)^{\frac{2}{m-1}}} \quad . \tag{4}$$

Ostatecznie należy zagregować otrzymany wynik. Analizowano różne podejścia do definicji ostatecznego wyniku. Najlepsze rezultaty przyniosła metoda, w której zapamiętywany jest prototyp każdego z liści, do którego został przyporządkowany obiekt w każdym z drzew kontekstowych. Następnie wartość przynależności jest wyliczana według wzoru (4) dla grupy prototypów będących reprezentacją liści wybranych w każdym z kontekstów i ostatecznie efektem jest przyporządkowanie obiektu według klasy wyliczonej w kontekście, dla którego obliczona wartość współczynnika przynależności jest największa.

3 Eksperymenty

Przedstawione zostaną przebieg i wyniki (Tabela 1 i 2) testów jakości badanego klasyfikatora wykonane przy użyciu aplikacji wykonanej w ramach pracy dyplomowej [7]. Aby otrzymane wyniki mogły zostać uznane za w pełni zasadne, porównanie wartości uzyskanych za pomocą różnych metod było przeprowadzone na identycznych danych treningowych i testowych. W utworzonej aplikacji dostępna jest opcja podziału zbioru danych na zbiór treningowy i testowy oraz zapis do pliku w celu późniejszego powtórzenia analizy na tej samej grupie danych. Do podziału danych zastosowano kroswalidację stratyfikowaną co oznacza, że zbiór treningowy i testowy zawiera zbiór obiektów o takiej samej proporcji liczności obiektów poszczególnych klas, jak w oryginalnym zbiorze danych. Może być to szczególnie istotne dla zbiorów danych, w których pojawia się znaczna dysproporcja pomiędzy liczbami obiektów należących do poszczególnych zmiennych decyzyjnych. Dodatkowo dane zostały podzielone pomiędzy zbiór treningowy i testowy w proporcji 80%, 20%.

Badania wykonano ze zmiennymi wartościami następujących parametrów:

- liczba klastrów *c* z wartościami od 2 do 15, ponieważ na większości analizowanych zbiorów danych nie było możliwe skonstruowanie drzew dla wartości przekraczających 15;
- liczba kontekstów *k* będącą liczbą nieparzystą z wartościami od 3 do 15 (w zbiorach, gdzie taki podział miał racjonalne zastosowanie);
- funkcja przynależności (gaussowska, trójkątna, trapezoidalna).

Aby zobrazować szerokie zastosowanie algorytmów klastrowych rozmytych drzew decyzyjnych oraz klastrowo-kontekstowych rozmytych drzew decyzyjnych, przedstawiono wyniki otrzymane dla zbiorów danych o różnej charakterystyce, przez co rozumiemy atrybuty oraz zmienną decyzyjną przyjmujące wartości zarówno dyskretne, jak i ciągłe. Badania przeprowadzone zostały na zbiorach: *Dermatology* oraz *Housing Data*. Są to powszechnie znane zbiory danych [8], często używane do testów jakości klasyfikatorów.

W zbiorze *Dermatology* zawarte są obserwacje 366 obiektów, z których każdy opisano zbiorem 34 cech, gdzie wszystkie cechy przyjmują wartości całkowite. Zmienna decyzyjna została odwzorowana wartościami od 1 do 6. Porównanie najlepszych wyników dla obydwu metod przedstawiono w Tabeli 1.

Tab. 1. Podsumowanie wyników otrzymanych dla obydwu metod

Tab. 1. Summary of the results obtained for both methods

Metoda i parametry	**Średni błąd zbioru treningowego (%)**	**Średni błąd zbioru testowego (%)**
Klastrowa, c = 2	2,72	2,78
Klastrowa, c = 4	1,36	1,39
Klastrowo-kontekstowa, k=3, c=2, f. gaussowska	**0,68**	**0**
Klastrowo-kontekstowa, k=3, c=2, f. trójkątna	**0,68**	2,78

Zastosowanie metody klastrowo-kontekstowych rozmytych drzew decyzyjnych umożliwiło otrzymanie lepszych wyników niż po zastosowaniu metody klastrowych rozmytych drzew decyzyjnych. Metoda ta, pomimo lepszych wyników nie zredukowała rozmiarów otrzymanych rozmytych drzew decyzyjnych.

Zbiór danych *housing.data* zawiera 506 wzorców. Każdy wzorzec opisany jest za pomocą 14 atrybutów (13 cech ciągłych, 1 boolowska) i zmienną decyzyjną, będące wartościami rzeczywistymi. Podsumowanie najlepszych wyników dla obydwu metod przedstawiono w tabeli 2.

Tab. 2. Podsumowanie wyników otrzymanych dla obydwu metod

Tab. 2. Summary of the results obtained for both methods

Metoda i parametry	**Średni błąd dla zbioru treningowego**	**Średni błąd dla zbioru testowego**
Klastrowa, c = 7	2,2078	3,25
Klastrowa, c = 5	1,8048	**2,3949**
Klastrowa, c = 3	1,6331	3,275
Klastrowo-kontekstowa, k=5, c=3, f. gaussowska	**0,2596**	3,0604
Klastrowo-kontekstowa, k=3, c=2, f. trójkątna	1,0868	2,9014

Drzewa klastrowo-kontekstowe są mniej złożone od drzewa otrzymanego metodą klastrową dla najlepszego uzyskanego wyniku.

Należy zauważyć, że najlepsze jakościowo wyniki klasyfikacji na większości zbiorów danych udało się uzyskać z wykorzystaniem gaussowskiej oraz trójkątnej funkcji przynależności, zależnie od wybranego zbioru danych. Widać, że najkorzystniejsze otrzymane wyniki uzyskano z zastosowaniem wykresów kontekstów dostosowanych odpowiednio do aktualnego zbioru danych co oznacza, że nie zawsze sprawdzała się taka sama konfiguracja. Dodatkowo zwiększenie liczby kontekstów oraz liczby klastrów w kontekście nie gwarantowało otrzymania lepszych wyników na każdym zbiorze danych.

Ostatecznie zastosowanie metody drzew klastrowo-kontekstowych pozwoliło uzyskać na większości zbiorów danych porównywalne lub lepsze wyniki klasyfikacji niż metoda klastrowych rozmytych drzew decyzyjnych.

4 Wnioski końcowe

Podsumowując, badania przeprowadzone na różnorodnych zbiorach danych pokazują, iż algorytm klastrowo-kontekstowych rozmytych drzew decyzyjnych umożliwia uzyskanie co najmniej podobnych lub lepszych wyników klasyfikacji w porównaniu do metody klastrowych rozmytych drzew decyzyjnych. Szczegółowe obserwacje zostały przedstawione przy omawianiu wyników dla każdego z analizowanych zbiorów danych. Główną cechą klastrowych rozmytych drzew decyzyjnych jest w większości przypadków mniejsza różnica w jakości klasyfikacji pomiędzy danymi treningowymi i testowymi. Ponadto widać, że algorytm klastrowo-kontekstowych rozmytych drzew decyzyjnych może być uważany za dobre rozwiązanie zarówno dla zbiorów danych o zmiennych ciągłych, jak i dyskretnych, co jest jego dużą zaletą. Analogicznie

do algorytmu klastrowych rozmytych drzew decyzyjnych zastosowanie grupowania rozmytego do dyskretyzacji atrybutów ciągłych umożliwia konstrukcję mniej złożonych drzew niż w przypadku standardowych metod budowy drzew decyzyjnych.

Literatura

1. Rutkowski L.: *Metody i techniki sztucznej inteligencji*, Wydawnictwo Naukowe PWN, 2006
2. Quinlann J.R.: *Induction of decision trees* , Machine Learning 1, 1986, 81-106
3. Pedrycz W., Sosnowski Z. A.: *C-Fuzzy Decision Trees*, IEEE Transactions on Systems, Man and Cybernetics, Part C, Vol. 35, No 4, 2005, p. 498-511
4. Cichosz P.: *Systemy uczące się*, Wydawnictwo Naukowo-Techniczne, Warszawa 2000
5. Pedrycz W., Sosnowski Z. A.: *Designing decision trees with the use of fuzzy granulation*, IEEE Transactions on Systems, Man, and Cybernetics – Part A, vol. 30, 2000,151-159
6. Sosnowski Z.A.: *Contex-based fuzzy cluster-oriented decision trees*, The European Simulation & Modelling Conference ESM'2007, October 22-24, 2007, St. Julians, Malta, p.326-328
7. Romanowicz M.K.: *Badanie metod hierarchicznego grupowania rozmytego*, praca magisterska, Politechnika Białostocka, Wydział Informatyki, 2011
8. 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 klastrowo-kontekstowych rozmytych drzewach decyzyjnych oraz zbadano jego wydajność na standardowych zestawach danych: *Dermatology* i *Housing Data Sets*. Wyniki symulacji pokazują, że przedstawiony klasyfikator daje zadowalające wskaźniki klasyfikacji.

Słowa kluczowe: klasyfikator rozmyty, rozmyte drzewa decyzyjne, rozmyte grupowanie kontekstowe

Decision rules with fuzzy granulation of knowledge

Summary

In this paper, we present the architecture of fuzzy classifier based on context fuzzy cluster-oriented decision trees and examine its performance on *Dermatology* and *Housing* data sets. Simulation results show that the presented classifier has a satisfactory classification rate.

Keywords: fuzzy classifier; fuzzy decision trees, context based FCM

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