AKADEMIA GÓRNICZO-HUTNICZA
IM. STANISŁAWA STASZICA

WYDZIAŁ ELEKTROTECHNIKI, AUTOMATYKI, INFORMATYKI I ELEKTRONIKI
KATEDRA INFORMATYKI

AUTOREFERAT ROZPRAWY DOKTORSKIEJ

BADANIA NAD NOWYMI ALGORYTMAMI GENEROWANIA DRZEW DECYZJI

MGR INŻ. MAKSYMILIAN KNAP

Promotor:
Prof. zw. dr hab. inż. Zdzisław S. Hippe
Wyższa Szkoła Informatyki i Zarządzania
w Rzeszowie

Kraków, 2009

W procesie odkrywania wiedzy i generowania modeli uczenia[1], ogromną rolę odgrywa tworzenie quasi-optymalnych drzew decyzji na podstawie zbiorów danych (reprezentujących informacje o zbiorze analizowanych obiektów). Wynika to z faktu, że drzewa te są niezwykle ważnym obiektem badań, zarówno praktycznych jak i teoretycznych, m. in. z uwagi na występowanie nietrywialnych problemów podczas ich projektowania oraz stosowania. Wydaje się także, iż algorytmy tworzenia drzew decyzji mogą w pewnym stopniu odtworzyć sposób myślenia człowieka podczas podejmowania decyzji.

Drzewo decyzji jest jedną z alternatyw strukturalnej informacji atrybut-wartość, umożliwiającą podział zbioru obserwacji, a więc podział zbioru danych, na klasy lub kategorie. Drzewo decyzji składa się z węzłów decyzyjnych, węzłów terminalnych (liści) i z łączących je gałęzi. Węzeł decyzyjny, umiejscowiony na szczycie struktury drzewa, jest nazywany korzeniem. Ogólnie biorąc węzły decyzyjne określają testy, które należy przeprowadzić względem danej wartości atrybutu, z jedną gałęzią dla każdego wyniku testu, przy czym w szerszym kontekście węzeł decyzyjny może zawierać dowolną funkcyjną kombinację atrybutów. Natomiast węzły terminalne reprezentują określoną klasę (kategorię) obiektów, stanowiąc tzw. liście: jednorodne dla obiektów tej samej kategorii, lub niejednorodne – w przypadku obiektów reprezentujących różne kategorie. Z kolei gałęzie reprezentują wartości wyników testu przeprowadzonego w danym węźle.

Zasadniczym celem budowy drzew decyzji jest ustalenie, jakie strategiczne pytanie należy postawić na temat zmiennych zależnych (tzn. jakie kryterium należy zastosować w celu wyboru właściwego atrybutu opisującego), które spowoduje podział zbioru danych na bardziej homogeniczne skupiska danych, z możliwie małym błędem rzeczywistym. W zasadzie jest to najważniejsza część algorytmu indukcji drzew decyzji. Poprawny dobór kryterium gwarantuje czytelną i nieskomplikowaną strukturę drzewa decyzji. W praktyce wybór właściwego atrybutu, dokonywany jest poprzez porównanie wartości zastosowanego kryterium dla wszystkich atrybutów opisujących. W sytuacji, gdy analizie poddawane są zbiory o dużej liczności obiektów, opisanych wieloma atrybutami posiadającymi wiele różnych wartości, możemy mieć do czynienia ze znaczną złożonością obliczeniową, zmierzającą do ustalenia (przed obliczeniem wartości właściwego kryterium), np. liczności wystąpienia danej wartości badanego atrybutu dla wszystkich występujących – w analizowanym zbiorze – kategorii obiektów.

Konstruowanie drzew decyzji następuje poprzez rekurencyjny podział zbioru danych do momentu, aż każdy powstały w wyniku podziału podzbiór albo jest jednorodny (to znaczy zawiera elementy jednej tylko klasy decyzji), lub przeważająca część jego elementów należy do jednej katego-

1 Przedstawione w niniejszym referacie rozważania dotyczą uczenia nadzorowanego.

rii. Utworzone w ten sposób (formalnie) drzewo decyzji przedstawia tzw. model uczenia, objaśniający struktury wiedzy, ukrytej w analizowanym zbiorze danych. Zbiór ten jest nazywany zbiorem (lub ciągiem) uczącym, natomiast jakość otrzymanego drzewa sprawdza się analizując (tzn. klasyfikując) przypadki zawarte w odrębnym zbiorze danych, zwanym zbiorem (lub ciągiem) testującym. Obydwa wspomniane zbiory – uczący i testujący – są de facto specjalnym opisem badanych obiektów, który może być zestawiony w postaci tablicy decyzji, w której wiersze reprezentują kolejne obiekty, zaś kolumny odpowiadają wybranym cechom tych obiektów. Informacja o przynależności danego obiektu do określonej kategorii (klasy decyzji) jest zapisana w ostatniej (skrajnej prawej) kolumnie tablicy. Ten sposób reprezentacji danych znany jest również w literaturze pod nazwą „danych typu 2A".

Zasadniczym zatem celem przeprowadzonych badań była próba opracowania nowych algorytmów generowania drzew decyzji. Pierwszy z nich (o nazwie **TVR**, **T**ree-**V**ia-**R**ule), tworzący drzewo decyzji z uprzednio wygenerowanych quasi-optymalnych reguł składniowych. Natomiast drugi algorytm (o nazwie **VCF**, **V**aried-**C**onfidence-**F**actor), wykorzystuje podczas generowania drzewa decyzji informację o istotności w procesie klasyfikacji poszczególnych atrybutów opisujących, pobraną z generowanej w tle – dla analizowanych danych – sieci przekonań Bayesa. W korzeniu tak generowanego drzewa, umieszczony zostaje atrybut opisujący, ujawniający największy wpływ marginalnego prawdopodobieństwa na atrybut decyzyjny.

W dostępnej literaturze dotyczącej odkrywania wiedzy w danych coraz większą uwagę poświęca się generowaniu modeli uczenia w warunkach niepewności, tzn. w przypadku, gdy analizowane dane zawierają przypadki sprzeczne. Z różnych metod przezwyciężania tej trudności, obecnie najczęściej wymienia się zastosowanie elementów teorii zbiorów przybliżonych (ang. *rough sets*). Wykorzystanie wspomnianych elementów podczas projektowania algorytmu **TVR**, pozwoliło – poprzez generowanie dolnego i górnego przybliżenia analizowanych danych – na przetwarzanie zbiorów uczących zawierających przypadki sprzeczne, dając w rezultacie drzewa *pewne* (dla dolnego przybliżenia) oraz *możliwe* dla górnego przybliżenia analizowanych danych. Wspomniane algorytmy zostały zaimplementowane w specjalnie opracowanym systemie analizy danych ***Tree*SEEKER**, składającym się z czerech podstawowych, realizujących odrębne funkcje, powiązanych ze sobą modułów programowych. Są to:

1. Moduł preprocesora, aktywowany na etapie wczytywania zbioru analizowanych danych, ujawniający błędy w nim występujące. Dodatkowo, na podstawie obliczanego iloczynu szacowanej liczby przypadków oraz liczby atrybutów opisujących, moduł ocenia wielkość wczytywanego zbioru. W przypadku gdy wynik wspomnianego iloczynu jest większy od 2000, moduł pozwala

na utworzenie reprezentacji oryginalnego zbioru, zawierającej proporcjonalną liczbę przypadków ze wszystkich kategorii występujących w zbiorze źródłowym.

2. Moduł budowy modelu uczenia – tzn. drzew decyzji – wraz z zaimplementowanymi algorytmami:
   a) algorytmem budowy drzew decyzji w oparciu o współczynniki Czerwińskiego, przystosowanym do zastosowań ekonomicznych,
   b) znanym algorytmem J.R. Quinlana **ID3/C4.5** – uznanym za wzorcowy dla opracowanego systemu, wykorzystywanym m. inn. do porównywania wyników otrzymywanych za pomocą algorytmów opracowanych w niniejszej dysertacji,
   c) nowo-opracowanymi algorytmami **TVR** oraz **VCF**.
3. Moduł testowania oraz prezentacji wyników, zawierający trzy odrębne bloki informacyjne: (**i**) listę badanych obiektów ze zbioru testującego; (**ii**) nazwę algorytmu za pomocą którego utworzono drzewo decyzji i przeprowadzono proces klasyfikacji, liczbę przypadków ze zbioru testującego, liczbę przypadków poprawnie i błędnie sklasyfikowanych oraz obliczony błąd klasyfikacji; (**iii**) tablicę rozproszenia, zawierającą liczby przypadków przypisane do danych kategorii. Jej przekątna zawiera liczby poprawnie sklasyfikowanych przypadków; powyżej przekątnej umieszczono liczby przypadków nazywane – w odniesieniu do danych medycznych – fałszywie dodatnimi, tzn. wskazującymi na bardziej groźny stan schorzenia niż jest w rzeczywistości. Natomiast przypadki fałszywie ujemne (poniżej przekątnej) wskazują mniej groźny stan schorzenia niż jest w rzeczywistości, co może prowadzić np. do zaprzestania leczenia i niezamierzonego pogorszenia stanu pacjenta.

Ostatnim elementem pracy były klasyfikacyjne badania porównawcze opracowanych algorytmów. Polegały one na porównywaniu wartości błędu klasyfikacji (uzyskiwanego dla poszczególnych drzew decyzji), określającego liczbę błędnie sklasyfikowanych obiektów ze zbioru testującego. Dodatkowo, do oceny jakości modeli uczenia, wykorzystano kryterium średniej liczby pytań (wzór 1). Porównanie tej wartości dla dwu (lub więcej) systemów identyfikujących ten sam zbiór przypadków, umożliwia dokonanie oceny rozpatrywanych systemów klasyfikacyjnych (najmniejsza wartość średniej liczby pytań określa drzewo quasi-optymalne).

$$E(S) = \sum_{i=1}^{n} S(c_i)\, p_i \qquad (1)$$

gdzie: $S(c_i)$ *jest liczbą pytań, które należy zadać aby zidentyfikować alternatywę* c_i*, natomiast* p_i *jest prawdopodobieństwem tej alternatywy.*

Opracowany system informatyczny poddano szczegółowym badaniom, wykorzystując bazy informacyjne zaczerpnięte z repozytorium Uniwersytetu Kalifornijskiego w Irvine, powszechnie używane w badaniach z dziedziny uczenia maszynowego. Badane bazy informacyjne były zróżnicowane pod względem rodzaju atrybutów oraz ich liczebności[2]. Zasadnicze badania zostały przeprowadzone w procesie analizy bazy informacyjnej znamion melanocytowych skóry, zawierającej 522 zweryfikowanych histologicznie przypadków, należących do jednego z czterech rodzajów wspomnianych znamion: (**i**) <*znamię łagodne*>, (**ii**) <*znamię błękitne*>, (**iii**) <*znamię podejrzane*> oraz (**iv**) <*znamię złośliwe*>. Atrybuty opisujące przypadki (w sumie 14-cie) ujęte w analizowanej bazie informacyjnej, mogą być formalnie przypisane do czterech grup zmiennych, tworzących znaną w dermatologii regułę **ABCD** {**A**symmetry, **B**order, **C**olor, **D**iversity of structures}. W regule tej (w odniesieniu do polskiej wersji językowej), **A** <*Asymetria*> – oznacza asymetrię znamienia, **B** <*Brzeg*> – informuje o charakterze obrzeża znamienia, **C** <*Kolor*> – zawiera informację o liczbie i rodzaju barw znamienia, natomiast **D** <*Struktura*> – mówi o liczbie i charakterze zróżnicowania dopuszczalnych struktur znamienia. Analizowana bazy zawierała dodatkowo atrybut **TDS** (ang. **T**otal **D**ermatoscopy **S**core), obliczany zgodnie z zasadą konstruktywnej indukcji cech na podstawie wartości atrybutów pierwotnych. Metodologia zrealizowanych badań polegała na: wygenerowaniu drzew decyzji wszystkimi dostępnymi algorytmami na wydzielonym zbiorze uczącym (zawierającym 348 przypadki) oraz przeprowadzeniu procesu klasyfikacji zbioru testującego zawierającego 174 przypadków. Zbiorcze wyniki badań zaprezentowane zostały w Tablicy 1.

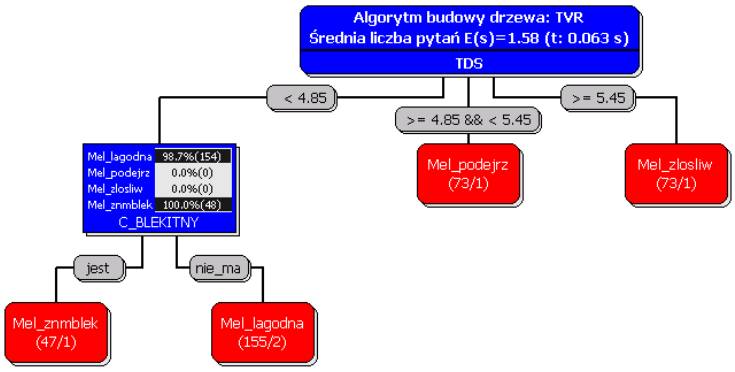
Tablica 1. Wyniki badań algorytmów generowania drzew decyzji, zaimplementowanych w systemie ***Tree*SEEKER** dla zbioru znamion melanocytowych skóry

Badane parametry	Algorytmy generowania drzew decyzji			
	wsp. Czerwińskiego	**ID3/C4.5**	**TVR**	**VCF**
Średnia liczba pytań	2,74	2,74	1,58	2,74
Czas generowania [ms][3]	47	62	63	47
Atrybut w korzeniu	TDS	TDS	TDS	TDS
Błąd klasyfikacji [%]	3,28	3,28	1,37	3,28

2 M. inn. były to zbiory: Hvote, Lenses, Cleveland, Autos, Bupa oraz Glass.

3 Podane czasy generowania poszczególnych drzew otrzymano na komputerze PC z procesorem Intel Core Duo, taktowanym zegarem 1,66GHz wyposażonym w 2GB pamięci operacyjnej oraz działającym pod kontrolą systemu operacyjnego Microsoft Windows XP.

Drzewem optymalnym okazało się drzewo wygenerowane przez nowo-opracowany algorytm **TVR** (jego struktura przedstawiona została na Rys. 1). Wartość średniej liczby pytań wyniosła 1,58 (dla porównania 2,74 dla pozostałych drzew), dodatkowo drzewo to cechuje się najlepszą skutecznością klasyfikacji, przy czym błąd klasyfikacji zbioru testującego wyniósł zaledwie 1,37%. Również wyniki uzyskane dla algorytmu **VCF** były zadowalające, ich wartość była porównywalna do wartości uzyskanych dla algorytmu **ID3/C4.5**.



*Rysunek 1. Optymalne drzewo decyzji wygenerowane przez nowo-opracowany algorytm **TVR** dla zbioru znamion melanocytowych skóry. W węźle początkowym (korzeniu drzewa) jest podana informacja o rodzaju zastosowanego algorytmu indukcji drzewa (tutaj algorytm **TVR**). W drugim wierszu etykiety tego węzła podano obliczoną (przez system) wartość średniej liczby pytań E(S) = 1,58 oraz czas generowania w [s]. Natomiast informacja „TDS" (trzeci wiersz etykiety) jest nazwą atrybutu opisującego, wybranego do testu w korzeniu. Atrybut ten ma trzy wartości: < <4,85 >, < >=4,85 && <5,45 > lub < >=5,45 >, zapoczątkowujące trzy ścieżki (gałęzie). Węzły końcowe (liście) zawierają, oprócz informacji o etykiecie rozpoznawanej klasy, dane o liczbie przypadków zlokalizowanych w danym węźle terminalnym. Przykładowo, w prawym skrajnym węźle zapis 73/1 wskazuje, że zbiór przypadków w tym węźle nie jest zbiorem jednorodnym i zawiera 73 obiekty, tj. 72 przypadki należące do <Mel_zlosliw> i 1 przypadek należący do innej kategorii znamienia*

Uzyskane wyniki badań wskazują, że opracowane w ramach niniejszej rozprawy nowe algorytmy generowania quasi-optymalnych drzew decyzji, generują drzewa cechujące się wysoką skutecznością klasyfikacji (porównywalną a często lepszą od skuteczności obserwowanej dla standardowego algorytmu **ID3/C4.5**), nie powodując odczuwalnego zwiększenia czasu generowania drzewa. Należy również podkreślić, że algorytm **TVR** umożliwia przetwarzanie przypadków sprzecznych (generując drzewo *pewne* oraz drzewo *możliwe*) i stanowi w ten sposób unikalne i niezastąpione narzędzie do interpretacji różnych danych tego typu, m. inn. danych medycznych. Przeprowadzone badania potwierdzają zasadniczą tezę rozprawy:

Przetwarzanie danych typu 2A przy pomocy sieci przekonań Bayesa lub z zastosowaniem rekurencyjnego algorytmu pokrycia do generowania reguł składniowych, przypuszczalnie może być podsta-

wą nowej metodologii budowy drzew decyzji, o skuteczności klasyfikacji porównywalnej z klasyfikacją takich algorytmów jak np. ID3/C4.5. Jednocześnie, zastosowanie wymienionych koncepcji w połączeniu z teorią zbiorów przybliżonych Pawlaka, umożliwi traktowanie przypadków sprzecznych, generując drzewa pewne lub drzewa możliwe.

Przedstawione w referowanej rozprawie drzewo decyzji – dotyczące klasyfikacji znamion melanocytowych skóry (Rys. 1) – może być z powodzeniem zastosowane przez lekarzy dermatologów, a także lekarzy pierwszego kontaktu, do precyzyjnego rozpoznawania-identyfikacji badanego schorzenia (wyniki prezentowanych badań zostały wykorzystane w opracowanym internetowym systemie diagnozowania znamion melanocytowych skóry, wspomagającym diagnozowanie wspomnianych znamion na podstawie cech opisujących badane znamię, dostępnym pod adresem www.melanoma.pl).