Investigationes Linguisticae, vol. XI, Poznań, December 2004

Cechy dystynktywne

Andrzej Pluciński

Zakład Fonetyki
Instytut Językoznawstwa Uniwersytet im. Adama Mickiewicza w Poznaniu
ul. Międzychodzka 5, 60-371 Poznań, Poland

apl@amu.edu.pl

Abstract

Three methods for distinctive (independent) features determination were designed and implemented in computer programs. Two of them were based on the assumption that the maximum number of the dependent variable's values cannot exceed the number of the points of the variable space. Using the third method, some functional relations between *a priori* chosen dependent features and the sets of *a priori* chosen independent features were studied. The main purpose of another method was to determine the minimum sets of features necessary to distinguish individual objects in a given set. It allowed to evaluate features' "load" (how often they were used in the task of object's distinguishing). In this way, any complex functional relations of any number of categorical features which can take any number of values can be determined. Possible applications vary from optimizing phones' description in the articulatory space, to the optimal selection of graphical features of different alphabets' signs, and database normalization.

2. Wstęp

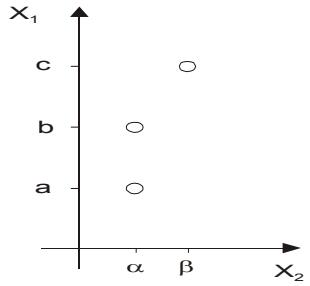
Każdy fizyczny obiekt charakteryzuje pewna realizacja zespołu przysługujących mu cech wyróżniająca go spośród innych obiektów zbioru. Do rozróżnienia obiektów wystarczy na ogół tylko podzbiór ich cech. Nie wszystkie cechy są niezależne. Cechy zależne nie wnoszą nic nowego do klasyfikacji. Cechy niezależne nazywa się „dystynktywnymi", a cechy zależne „redundantnymi". Przedmiotem tej pracy jest wybór cech niezależnych na podstawie oceny równomierności rozkładów obiektów w poszczególnych klasach i liczebności klas wyróżnianych poprzez poszczególne cechy albo na podstawie analizy zależności funkcyjnych. Zakładamy, że cechy przyjmują wartości nominalne.

Cechy zależne, to cechy w pełni skorelowane z cechami niezależnymi. Problem wyodrębnienia cech niezależnych został pomyślnie rozwiązany dla zmiennych ilościowych (np. analiza korelacji, czy regresji). Stosowane są tam modele losowe, w których postępowanie opiera się na założeniu, że obserwowane wartości są bliskie prawdziwym z prawdopodobieństwem odwrotnie proporcjonalnym do rozbieżności. Przyjmuje się, że błędy obserwacji są skutkiem niedoskonałości metody pomiaru.

Modele stosowane dla zmiennych ilościowych nie mogą być stosowane dla zmiennych nominalnych. Ich wartości nie wyrażają natężenia cechy, są nieuporządkowane, a więc równoodległe – nie można więc stosować powyższego modelu losowego (ze względu na założenie o prawdopodobieństwie rozbieżności, które wymaga uszeregowania wartości zmiennej). Zmienne skorelowane można tu rugować sprawdzając, czy nie pozostają one w relacji funkcyjnej z innymi. Można też stosować podejścia opierające się na założeniu, że zmienna zależna przyjmuje co najwyżej tyle wartości, ile przyjmuje ich zmienna niezależna. My stosujemy obydwa te podejścia.

Opracowano trzy metody określania zbioru cech dystynktywnych. Dwie bazujące na algorytmach zachłannych i trzecią, pozbawioną tej wady, opartą na badaniu relacji funkcyjnej pomiędzy

cechami a zadanymi odgórnie przestrzeniami cech niezależnych. Znaleziono nową formułę współczynnika korelacji, który można stosować do oceny stopnia zależności pomiędzy zmiennymi jakościowymi. Jedna z wymienionych metod pozwala także określić zbiór cech minimalnych dla



Rys. 1. *Cecha x_2 nie może objaśnić wartości cechy x_1, bo przyjmuje od niej mniej wartości, co skutkuje przyporządkowaniem postaci jeden do wielu (np.* $\alpha \rightarrow \{a, b\}$*)*

każdego obiektu, który pozwala odróżnić go od innych obiektów z danego zbioru, co z kolei pozwala ocenić tzw. „obciążenie cech" w zadaniu rozróżniania obiektów z danego zbioru.

Problem określenia cech dystynktywnych podjął Lapis (2001). Zaproponował on algorytm heurystyczny oparty na podziałach dychotomicznych. Zdaniem jego dokonując wszystkich takich podziałów zbiorów wartości cech i badając współwystępowanie realizacji otrzymanych w ten sposób cech dwuwartościowych można dojść do cech dystynktywnych. Rozważania szczegółowe ogranicza jednak tylko do cech dwuwartościowych. Z pewnością lepszym rozwiązaniem byłoby zastosowanie techniki CART (*Correlation And Regression Trees*) opracowanej przez Breimana (Breiman, 1984, Salford Systems, 2003)[1]. Jest to jednak tzw. algorytm zachłanny, lokalnie optymalny (por. Cichosz, 2000) nie koniecznie wykrywający wszystkie współzależności.

Opisane metody wykrywania cech dystynktywnych mogą znaleźć zastosowanie wszędzie tam, gdzie mamy do czynienia z cechami jakościowymi, a więc np. w badaniach językoznawczych, gdzie głównie takie cechy występują. Opracowano je z myślą o zastosowaniu do optymalnego opisu głosek w przestrzeni artykulacyjnej oraz do zwięzłego określania cech graficznych znaków z różnych alfabetów. Ze względu na zdolność do wykrywania dowolnych relacji funkcyjnych mogą znaleźć zastosowanie w badaniach struktur baz danych – w zadaniu normalizacji trzeciego stopnia (por. np. Roman, 1999).

3. Metody wyboru cech dystynktywnych

Znaleźliśmy trzy metody wyboru cech dystynktywnych. Są to:

- metoda klasyfikująca obiekty według wartości ich cech – cechy dystynktywne to te, które zostały w tej klasyfikacji użyte,
- metoda określająca minimalne zbiory cech koniecznych do wyodrębnienia poszczególnych obiektów – suma tych zbiorów daje zbiór cech dystynktywnych,
- metoda sprawdzająca, czy pewna cecha nie jest funkcją innych; jeśli tak, to jest skorelowana i na pewno nie jest dystynktywną.

Pierwsza i druga metoda nie prowadzi wprost do wytkniętego celu. Otrzymywany tu zbiór cech dystynktywnych jest jakby efektem ubocznym postępowania. Pierwsza metoda dostarcza informacji o tym, czy jest wystarczająco dużo cech, aby rozróżnić wszystkie obiekty (pokazując nie rozbite zbiory obiektów), pokazuje z dużą pewnością wszystkie cechy skorelowane oraz konstruuje optymalne, deterministyczne drzewo decyzyjne (por. Dutoit 1997), które może być użyte dla celów identyfikacji. Druga metoda również ruguje cechy skorelowane, a oprócz tego rozwiązuje także inne, samodzielne zadanie. Ostatnia metoda prowadzi wprost do wykonania zadania. Metoda ta pozwala z całą pewnością wskazać, które cechy są wzajemnie skorelowane, a nie tylko te, które nie są niezależne.

[1] Przykłady zastosowań tej techniki można znaleźć m. in. w: Wang M., Q., Hirschberg J. (1992).

Uzasadnieniem postępowania w dwóch pierwszych metodach jest spostrzeżenie, że cecha zależna może przyjmować co najwyżej tyle wartości, ile przyjmuje ich cecha niezależna (rys.1). Ogólniej, cecha zależna przyjmuje co najwyżej tyle wartości, ile jest punktów w przestrzeni cech niezależnych (iloczyn liczb wartości przyjmowanych przez cechy składające się na przestrzeń cech niezależnych). Wynika stąd sposób na określenie siły dyskryminacji cech, a mianowicie za pomocą liczby wartości, jaką cecha w badanym zbiorze obiektów przyjmuje i za pomocą entropii rozkładu obiektów w funkcji wartości cechy. Sortowanie cech według liczby wartości i według entropii pozwala odrzucić cechy skorelowane w procesie klasyfikacji obiektów i w procesie poszukiwania minimalnych podzbiorów izolujących.

4. Wybór cech dystynktywnych metodą klasyfikacji

4.1. Zasada

Do zbioru cech dystynktywnych dochodzimy poprzez hierarchiczny podział zbioru obiektów na klasy według wartości ich cech. W postępowaniu z tym związanym wybieramy cechę o największej sile dyskryminacji w dzielonym podzbiorze. Ponieważ raz użyta cecha nie może być użyta w podziałach potomnych[2], więc w ten sposób wyselekcjonujemy zbiór cech dystynktywnych. Wybrane zostaną cechy o największej sile dyskryminacji, a cechy redundantne zostaną odrzucone. Odrzucenie każdej cechy redundantnej wynika stąd, że cecha taka nie będzie dzielić dalej podzbiorów wyróżnionych wartościami cechy lub cech, w stosunku do której lub których jest ona redundantna. Ponieważ będzie wyróżniać co najwyżej tyle samo klas obiektów i będzie mieć co najwyżej taką samą entropię, więc będzie brana pod uwagę w drugiej kolejności – a więc nie dojdzie w ogóle do głosu.

4.2. Siła dyskryminacji

Szukając miary siły dyskryminacji brano pod uwagę liczbę wyróżnionych klas i równomierność podziału zbioru obiektów. Ponieważ cecha zależna przyjmuje co najwyżej tyle wartości, ile przyjmuje ich cecha niezależna, więc kryterium liczby wartości powinno być najpierw brane pod uwagę. W razie, gdy istnieje kilka cech o tej samej liczbie realizowanych wartości, to należy wybierać tę, która ma większą entropię, ponieważ jest to cecha o bardziej równomiernym rozkładzie wartości (patrz uzasadnienie w punkcie 2.4).

Entropia histogramu wartości cechy, to

$$H = -\sum_{i=1}^{N} p_i \log_2 p_i = -\sum_{i=1}^{N} \frac{r_i}{R} \log_2 \frac{r_i}{R} = -\frac{1}{R} \sum_{i=1}^{N} r_i \left(\log_2 r_i - \log_2 R \right) \quad i=1,\ldots,N,$$

gdzie N oznacza liczbę wartości, jakie cecha przyjmuje w klasyfikowanych obiektach, r_i liczbę obiektów przyjmujących i-tą wartość cechy, R całkowitą liczbę obiektów, a p_i prawdopodobieństwo, że jakiś obiekt z klasyfikowanego zbioru znajdzie się i-tej klasie. Zakładamy, że bierzemy pod uwagę tylko te wartości cechy[3], które przyjmuje ona w klasyfikowanych obiektach, stąd wymóg, aby $p_i>0$. Jeśli rozkład jest równomierny, to $r_i = R/N$, stąd $p_i = R/\left(N \cdot R\right) = 1/N = const$. Entropia przyjmie wtedy wartość największą i równą $\log_2 N$, bo

$$H = H_{\max} = -\sum_{i=1}^{N} \frac{1}{N} \log_2 \frac{1}{N} = \log_2 N .$$

[2] Bo każdy z wyróżnionych podzbiorów składa się z elementów mających tę samą wartość użytej już cechy.

[3] Na obiekt realizowany przy więcej niż jednej wartości cechy można patrzeć jak na kilka różnych obiektów rozróżnianych wartościami tej cechy. Nie będzie to mieć żadnego wpływu na eliminację cech skorelowanych.

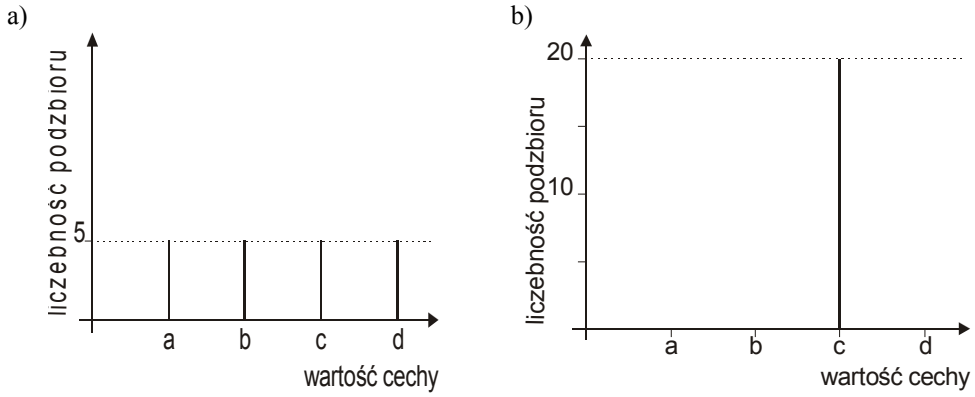
Jeśli zaś rozkład jest jednopunktowy, to

$$\underset{k\in\langle 1,\ldots,N\rangle}{\forall} p_i = \begin{cases} R/\left(N\cdot R\right)=1 & dla \quad i=k \\ 0 & dla \quad i\neq k \end{cases}$$

bo z założenia N=1. Wtedy

$$H = H_{\min} = \log_2 p_k = \log_2 1 = 0$$

(por. rys. 2). Entropia przyjmuje więc wartość zależną nie tylko od stopnia równomierności, lecz także i od liczby klas podziału.



Rys. 2. Zależność entropii od kształtu rozkładu: (a) przy rozkładzie równomiernym czteropunktowym *H=Hmax=log₂ 4=2,* (b) przy rozkładzie jednopunktowym *H=Hmin=log₂ 1=0*

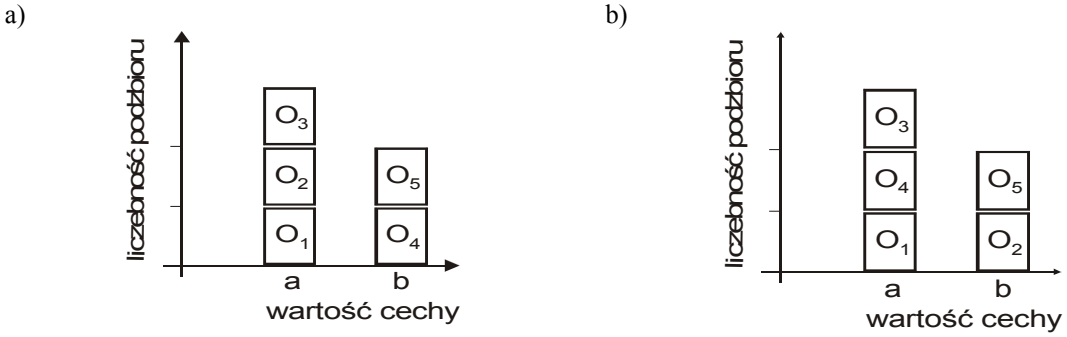
Entropia musi być obliczana na podstawie rozkładów o niezerowych wartościach chociażby dlatego, że log 0 nie istnieje. W sytuacji, gdy cecha może przyjąć więcej wartości niż jest obiektów, to w histogramie jej wartości pojawią się zerowe słupki. Przyjmujemy, że cechy mogą przyjmować nieograniczoną liczbę wartości, ale przeglądamy tylko zbiór zrealizowanych wartości.

Siła dyskryminacji definiowana za pomocą entropii stawia na równi liczbę podzbiorów podziału oraz równomierność rozkładu. W zadaniu określania zbioru cech dystynktywnych ważniejsza, jak powiedziano, będzie liczba klas podziału aniżeli jego równomierność.

Dalszym kryterium może być liczba wartości przyjmowanych przez cechę. Racjonalny wydaje się wybór cechy o najmniejszej liczbie wartości, jakie może ona w ogóle przyjąć[4]. Ostatecznym kryterium wyboru może być kolejność, w której wymieniono cechy obiektów.

[4] Cecha określona w dziedzinie liczb rzeczywistych może przyjąć nieskończoną liczbę wartości w każdym przedziale zmienności, będzie więc według tego kryterium uznana za gorszą w porównaniu z cechą określoną w dziedzinie liczb całkowitych.

Siła dyskryminacji, jako ocena punktowa kształtu rozkładu, nie może zawierać wielu informacji szczegółowych. Entropia np. nie rozróżni rozkładów, gdzie cechy dzielą zbiór obiektów na równoliczne, ale różne podzbiory (rys. 3).



Rys. 3. *Różne podziały o równej entropii*

4.3. Osłabienie zachłanności algorytmu

Decyzje o wyborach cech niezależnych zapadają na podstawie analiz rozkładów obiektów w podzbiorach kolejnych podziałów. Za każdym razem wybierana jest tylko jedna, „najsilniejsza" cecha i włączana do zbioru cech niezależnych. Może więc zdarzyć się, że cecha zależna może lokalnie, w podziałach potomnych, dorównać siłą dyskryminacji cesze niezależnej i w konsekwencji może dojść do włączenia jej do zbioru cech niezależnych. Aby ten efekt osłabić, przyjęliśmy że większą wagę mają cechy, które uzyskały większą siłę dyskryminacji w większych podzbiorach obiektów aniżeli aktualnie rozbijany. Przyjęliśmy więc dodatkowe kryteria wyboru cech oraz sortujemy podzbiory każdego podziału.

Spośród każdych dwóch cech o równej sile dyskryminacji wybieramy tę, która była już użyta we wcześniejszych podziałach – jej wybór nastąpił na podstawie większego zbioru obiektów, gdzie miała większą siłę dyskryminacji. Gdyby to nie wystarczyło, to bierzemy pod uwagę entropię globalną cech, obliczoną dla całego zbioru obiektów. Możliwość wcześniejszego użycia zależy jednak od kolejności wyboru podzbiorów potomnych do dalszej klasyfikacji. W pierwszej kolejności powinny być klasyfikowane najliczniejsze podzbiory, ponieważ można wtedy liczyć na to, że cecha niezależna, która mogłaby lokalnie nie przeważyć cechy zależnej zostanie użyta do klasyfikacji w większym zbiorze i tym samym zostanie zachowana jako cecha niezależna.

Podzbiory każdego podziału sortujemy wg wielkości. Pozwala to wybierać do dalszych podziałów najliczniejsze zbiory w pierwszej kolejności. Cechy wybrane wcześniej będą, zgodnie z powyższymi kryteriami, dominować nad pozostałymi.

4.4. Uzasadnienie stosowania entropii jako drugiego kryterium wyboru cech

Rozkład o większej entropii jest bardziej równomierny, zawiera więc mniej liczne zbiory, a takie łatwiej rozbić.

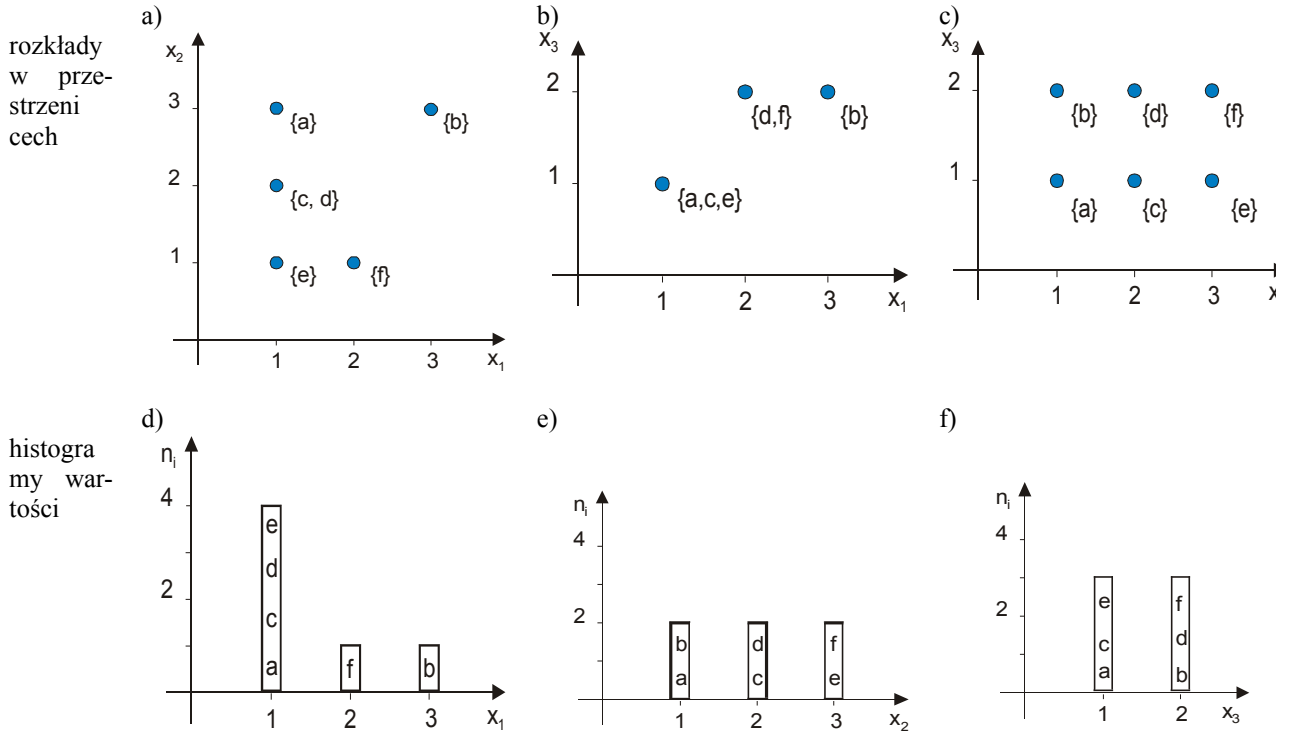
Przykład. Załóżmy, że mamy 6 obiektów i 3 cechy. Niech obiekty te będą realizacjami cech o wartościach wymienionych w tabeli 1.

Tabela 1. *Dane przykładowe*

cecha\obiekt	a	b	c	d	e	f	entropia
x_1	1	3	1	1	1	2	1,25
x_2	3	3	2	2	1	1	1,58
x_3	1	2	1	2	1	2	1,00

Rozkłady tych obiektów w różnych układach cech pokazano na rys. 4.

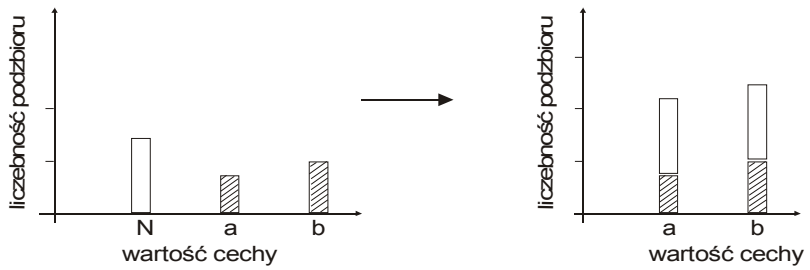
Widać, że zmienna x_1, o takiej samej liczbie realizowanych wartości jak zmienna x_2, ale o mniejszej entropii niż x_2, nie rozbija zbioru obiektów ani wespół ze zmienną x_2, ani wespół ze zmienną x_3 (rys. 4a i 4b). Cecha x_3 jako dwuwartościowa, nie jest w stanie rozbić zbioru czteroelementowego podzbioru wyróżnianego w wymiarze x_1, ale może rozbić dwuelementowe podzbiory wyróżniane w wymiarze cechy x_2 (rys. 4c). Okazuje się więc, że cecha x_1 jest zależna jednocześnie od cech x_1 i x_3.



Rys. 4. *Rozkłady obiektów w przestrzeni par cech i histogramy ich wartości*

4.5. Przysługiwanie bądź nie pewnej cechy

Pewne cechy mogą nie dotyczyć wszystkich obiektów, powstaje więc pytanie, jak określać siłę dyskryminacji w przypadkach, gdzie tylko część obiektów podzbioru jest określona w wymiarze pewnej cechy. O cechach, które nie są określone, mówi się, że „nie przysługują" obiektowi. Jeśli



Rys. 5. *Przekształcenie funkcji rozkładu uwzględniające istnienie cech nie przysługujących; N – wartość nieokreślona*

cecha nie przysługuje, to obiekt nie jest określony w jej wymiarze. Formalnie znaczy to, że może on się zrealizować przy każdej wartości owej nie przysługującej mu cechy. Zakładamy więc istnienie tylu możliwych realizacji tych obiektów, ile cecha nie przysługująca im może przyjąć

wartości. Funkcję rozkładu liczebności obiektów modyfikujemy tak, że do każdej liczebności zbioru elementów dodajemy jeszcze liczbę obiektów nieokreślonych. Jeśli liczbę tę oznaczymy symbolem c, to we wzorze na entropię symbole r_i oraz R należy zastąpić wyrażeniami r_i+c oraz $R+c$ odpowiednio (por. rys. 5). Jak łatwo zauważyć, obecność elementów nieokreślonych pomniejsza entropię. W rozwiązaniu technicznym posługujemy się chwytem polegającym na tym, że do zbioru wartości każdej cechy dodajemy jeszcze jedną wartość, którą umownie przypisujemy obiektom nieokreślonym. Ponadto do zbioru obiektów dodajemy jeden obiekt neutralny, nieokreślony w przyjętej przestrzeni cech, czyli przyjmujący tylko umowną wartość nieokreśloną w wymiarze każdej z nich. Obiekt neutralny umieszczamy jako pierwszy na liście, co powoduje, iż umowna wartość nieokreślona przyjmuje, przy przyjętej technice kodowania, numer 1. To z kolei sprawia, że wszystkie obiekty nieokreślone w wymiarze pewnej cechy będą zawsze wraz z elementem neutralnym zaliczane do zbioru nr 1, pozwalając w konsekwencji uprościć algorytm przetwarzania danych.

Algorytm

1. Uszeregować cechy według siły dyskryminacji w wybranym zbiorze[5] obiektów.
2. Wybrać cechę najsilniejszą lub, spośród jednakowo silnych, tę która była już użyta.
3. Podzielić zbiór obiektów na podzbiory według wartości tej cechy.
4. Posortować podzbiory malejąco według liczebności.
5. Wybrać pierwszy podzbiór podziału, który zawiera więcej niż jeden element, i iść do punktu 1. jeśli nie wyczerpano jeszcze zbioru cech poprzez użycie ich w podziałach nadrzędnych.
6. Jeśli osiągnięto podzbiór, którego dalej nie można już dzielić, to wybrać następny wieloelementowy podzbiór ostatniego podziału i iść do punktu 1.
7. Jeśli przeanalizowano wszystkie podzbiory bieżąco rozpatrywanego podziału, to wrócić do wcześniejszego podziału, wybrać nie podzielony jeszcze wieloelementowy podzbiór i iść do punktu 1.
8. Koniec.

Zakończenie analizy na wieloelementowym zbiorze obiektów będzie oznaczać, że wzięto pod uwagę za mało cech dystynktywnych.

4.6. Porównanie z techniką CART

Inspiracją dla nas przy tworzeniu opisywanej metody była technika CART. Podejście to poddaliśmy jednak głębokiej rewizji. Poniżej wskazujemy na różnice pomiędzy nimi.

W obydwu technikach w wyniku otrzymujemy drzewo klasyfikacji hierarchicznej obiektów według wartości cech.

W technice CART dzieli się zbiór wartości cechy na dwa podzbiory na wszelkie możliwe sposoby i bada rozkłady obiektów w wydzielonych tymi podziałami podzbiorach (obiektów). Następnie wybiera się najlepsze rozszczepienie. Najlepsze rozszczepienie to to, które maksymalizuje informację wzajemną pomiędzy stanem sprzed rozszczepienia a stanem po rozszczepieniu. Postępowanie to stosuje się kolejno wobec wszystkich cech i wybiera tę cechę, która spełnia najlepiej powyższe kryterium. Cechy zależne to te, które nie wezmą udziału w klasyfikacji obiektów. Widać, że postępowanie jest lokalnie optymalne (zachłanne) i nie koniecznie wyłączające wszystkie cechy zależne.

W naszym podejściu obiekty dzieli się od razu według wszystkich wartości cechy, co zmniejsza o kilka rzędów wielkości ilość potrzebnych obliczeń przy takich samych, jeśli nie lepszych wynikach końcowych. Cechy, jak to obszernie opisano, wybiera się na podstawie siły dyskryminacji. Łatwiej jest u nas osłabić zachłanność algorytmu i uzyskujemy lepsze, krótsze drzewo rozbioru.

[5] Można ograniczyć się tu do cech, które nie brały jeszcze udziału w podziałach nadrzędnych, bo te nie będą już dalej dzielić badanego zbioru.

Przewagą techniki CART jest zdolność do korekty błędów opisu obiektów na podstawie statystyk – podziały obiektów przerywa się przy pewnym progu informacji wzajemnej i – godząc się na ryzyko popełnienia błędu – wszystkim obiektom z końcowych podzbiorów przypisuje się te same wartości cech, które ma najwięcej obiektów w podzbiorze.

5. Zbiory minimalne

5.1. Wstęp

Postępowanie wykorzystujące klasyfikacje do określenia cech zbędnych nie daje odpowiedzi wprost na istotne dla praktyki pytanie o podzbiory cech wystarczających do rozróżnienia poszczególnych obiektów w ich przestrzeni. Liczba potrzebnych cech będzie zależeć od liczby i rodzaju obiektów.

Obiekty są różnie rozłożone w przestrzeni cech. Liczba cech wystarczająca dla ich identyfikacji może być różna. W przyjętym odgórnie zbiorze cech mogą wystąpić też cechy skorelowane z innymi. Te nie będą nic wnosić do możliwości klasyfikacji. Zadaniem tu rozwiązywanym jest określenie minimalnych podzbiorów cech koniecznych do identyfikacji wszystkich oraz każdego obiektu z osobna. Suma ich określi też podzbiór cech niezależnych, co oznacza, że wartości pozostałych cech dadzą się objaśnić wartościami cech z owego podzbioru cech. Będą to cechy skorelowane z jedną lub więcej cechami ze zbioru cech niezależnych.

5.2. Metoda

Postępowanie w tej metodzie polega na sprawdzaniu w przestrzeniach podzbiorów cech, czy istnieją w nich punkty realizowane w jednoelementowych zbiorach obiektów. Minimalnego zbioru cech poszukujemy więc sprawdzając:

- czy są obiekty o unikalnych wartościach pojedynczych cech, a następnie
- czy są obiekty o unikalnych wartościach w wymiarach par cech,
- jak wyżej, dla trójek, czwórek itd. aż do wyczerpania wszystkich ich kombinacji.

Kolejność wyboru cech nie jest obojętna. Aby wyodrębnić minimalny podzbiór cech (będą to cechy niezależne) należy wybierać cechy według siły dyskryminacji. Wtedy już na początku postępowania zostanie wyizolowanych dużo obiektów, które następnie należy wyłączyć z procedury przeglądania. Wybierając w pierwszej kolejności cechy o największej entropii globalnej, można liczyć na to, iż będą one zwyciężać we wszystkich podprzestrzeniach, w których będą brane pod uwagę. Ponieważ proces izolacji rozpoczynamy od przestrzeni jednowymiarowych, a kończymy na przestrzeni uwzględniającej wszystkie cechy i ponieważ obiekty wyizolowane są wyłączane ze zbioru, to obiekty te nie będą już określane w innych przestrzeniach niż za pierwszym razem. To oznacza, że użyty zbiór cech będzie z dużym prawdopodobieństwem najmniejszy z możliwych[6].

Zliczając, ile razy każda z cech była użyta do wydzielenia obiektów, można wskazać cechy skorelowane – będą to te, które nie zostały użyte ani razu.

6. Poszukiwanie cech skorelowanych

6.1. Wstęp

Wybór podzbioru zmiennych niezależnych w sytuacji, gdy mamy do czynienia z wieloma zmiennymi i wieloma obiektami nie jest łatwy, zwłaszcza gdy zachodzi jednoczesna zależność od wielu zmiennych. Przedstawiamy tu metodę bazującą na badaniu zależności funkcyjnej pomiędzy zbiorami wartości zmiennej, o której przypuszczamy, że jest zależna, a punktami w przestrzeniach podzbiorów innych cech, które traktujemy jako niezależne. Podajemy miarę skorelowania pewnej cechy z zespołem innych cech.

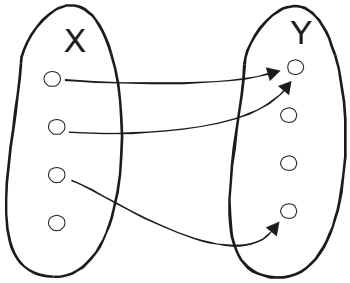
[6] Wypróbowaliśmy też podejście, w którym najpierw wybierano cechy wg liczby klas, a następnie, w ramach tych wyborów, cechy o największej entropii globalnej. W efekcie nie uzyskano jednak eliminacji cechy skorelowanej, jaką wskazała metoda klasyfikacji. W teście tym przewagę nad pewną inną uzyskała cecha o większej wyróżnianych przez się liczbie klas, ale o mniejszej entropii. W praktyce należy zatem próbować jednej i drugiej możliwości i wybierać tę, która daje mniejszy zbiór cech dystynktywnych.

6.2. Metoda

Korelacje zmiennych nominalnych można badać wychodząc z założenia, że cechy skorelowane tworzą odwzorowanie funkcyjne, tj. że zachodzi relacja

$$y = f(\vec{x}),$$

gdzie *y*, to cecha zależna, $\vec{x}$ - punkt w przestrzeni pewnego podzbioru cech niezależnych (objaśniających) $\vec{x} = (x_1 \ldots x_n)^T$. Wtedy każdemu punktowi $\vec{x}$ odpowiada tylko jeden punkt w wymiarze cechy *y* (lecz nie na odwrót!, por. rys. 6).



Rys. 6. *Przykład odwzorowania funkcyjnego*

Miarą skorelowania może zatem być współczynnik oparty na liczebnościach punktów z *X* przypisanych punktom w *Y*. Współczynnik ten należy skonstruować tak, aby przyjmował wartości z przedziału $\langle 0,1 \rangle$. Wymogi te spełnia

$$r_{X,Y} = \frac{N_X - 1}{N_Y - 1},$$

gdzie $N_X = \sum_{i:n_i>0} 1$, a $N_Y = \sum_i n_i$. Symbol n_i oznacza liczbę punktów w wymiarze *y* przypisanych *i*-temu punktowi w *X*, N_X oznacza liczbę punktów w *X*, dla których $n_i > 0$.

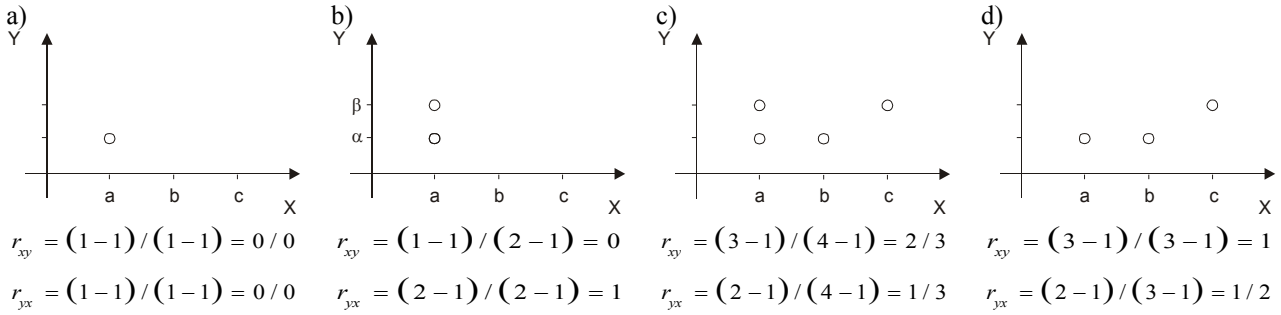
Skutkiem realizacji cech są pewne obiekty fizyczne, a więc korelacje możemy badać analizując cechy pewnego zbioru obiektów. Na podstawie wartości jednej lub kilku wybranych cech zbiór można podzielić na podzbiory. Powiększając liczbę branych pod uwagę cech otrzymane podzbiory można dalej dzielić według ich wartości itd. Każda dodatkowo wzięta pod uwagę cecha ten podział pogłębi pod warunkiem, że nie będzie ona skorelowana z cechami już uwzględnionymi. Cecha skorelowana, ze względu na zakładane w takim przypadku odwzorowanie funkcyjne, nie będzie różnicować elementów już wyróżnionych podzbiorów.

Każdy punkt w przestrzeni cech reprezentuje pewien podzbiór obiektów. Współczynnik korelacji cechy *y* z zespołem cech *X* można więc obliczać wg formuły

$$r_{XY} = \frac{\textit{liczba podzbiorów w przestrzeni } X - 1}{\textit{liczba podzbiorów w przestrzeni } X \times y - 1}.$$

Współczynnik ten osiąga wartość 1, gdy zachodzi relacja funkcyjna, wtedy $N_Y = N_X$. W pozostałych przypadkach $N_Y > N_X$. W skrajnym przypadku wszystkie punkty w *y* mogą być przypisane jednemu punktowi w *X*. Wtedy $N_X = 1$ i *r=0*. O korelacji nie można niczego powiedzieć, gdy $N_Y = 1$, wtedy *r=0/0* staje się symbolem nieokreślonym. Sytuacje te ilustruje rysunek 7.

Przykład pokazany na rys. 7. pokazuje, że relacja bycia w korelacji nie jest symetryczna, skutkiem czego również macierz korelacji nie będzie symetryczna.



Rys. 7. *Różne stopnie skorelowania cechy y z cechą x (zaznaczone punkty reprezentują dowolnie liczne podzbiory, a nie pojedyncze obiekty)*

Algorytm

1. Wybierz krotkę cech mających tworzyć przestrzeń zmiennych niezależnych.
2. Określ rozkład obiektów w wybranej przestrzeni cech.
3. Dokładaj, a następnie odejmuj po jednej z pozostałych cech i określaj rozkład obiektów w powstałej przestrzeni.
4. Na podstawie tych rozkładów (określonych przed i po dołożeniu cechy) obliczaj współczynnik korelacji cechy dodanej z cechami krotki.
5. Przejrzyj w ten sposób wszystkie kombinacje cech.

7. Zakończenie

Konkluzje. Wszystkie trzy opisane metody mogą wskazać cechy skorelowane, każda według zupełnie innego sposobu postępowania. Konfrontacja wyników będzie zatem sprawdzianem ich poprawności. Oprócz tego każda oferuje dodatkowo inne wyniki analiz.

Pierwsza metoda dostarcza także drzewa klasyfikacji hierarchicznej, które może być wykorzystane do wskazywania zbiorów obiektów o określonych cechach.

Druga metoda określa dla każdego obiektu minimalne zbiory cech niezbędnych do ich identyfikacji. Oblicza też obciążenie cech, tj. dla każdej cechy podaje liczbę obiektów, dla identyfikacji których jest ona niezbędna.

Trzecia metoda oblicza współczynniki korelacji. Cecha zależna to ta, która ma współczynnik korelacji równy 1. Duży współczynnik korelacji, ale mniejszy od 1, oznacza, że cecha typowana jako zależna wprawdzie taką nie jest, ale izoluje niewiele obiektów. Dodatkowa inspekcja tych właśnie obiektów może ujawnić błędy w ich opisie, a więc metoda ta może także pomóc w usuwaniu błędów opisu. Znane z literatury współczynniki korelacji, takie jak Spearmana czy Kendalla (por. np. Ferguson, 1997) nie mogą być tutaj zastosowane, ponieważ wymagają przypisania liczb poszczególnym wartościom cechy, to zaś oznacza uporządkowanie ich w skali odległości, co, jak powiedziano, nie może mieć miejsca w przypadku zmiennych nominalnych (wartości ich nie wyrażają natężenia cechy, wszystkie są więc równoodległe). Ponadto współczynniki te dotyczą tylko dwóch zmiennych, co byłoby niewystarczające w rozwiązywanym tu zadaniu.

Trzecia metoda, w odróżnieniu od poprzednich, pokazuje cechy od których pewna cecha jest zależna, a nie tylko że nie jest ona niezależna.

Implementacja. Program komputerowy implementujący przedstawione metody wyposażono w wiele opcji pozwalających w szerokim zakresie zmieniać warunki analiz, m.in. włączać i wyłączać z rozważań cechy obiektów, określać ich priorytet i warunki wyboru.

Testy. Implementacje komputerowe omawianych metod testowano między innymi różnymi zestawami danych generowanych losowo (z liczbą obiektów od 10 do 1000, liczbą cech od 3 do 20 przyjmujących 2 do 20 wartości). We wszystkich testach metody te sygnalizowały te same zbiory cech zależnych i niezależnych. Program przetestowano również za pomocą danych dotyczących

cech artykulacyjnych udostępnionych przez prof. dr hab. M. Steffen-Batogową. z Zakładu Fonetyki Instytutu Językoznawstwa UAM.

Analiza zbioru zawierającego 136 obiektów opisywanych dziewięcioma cechami o różnych liczbach wartości trwa ok. 8 s przy metodzie klasyfikacji, 20 s przy określaniu zbiorów minimalnych i 6 s przy analizie korelacji (PC, procesor Athlon, zegar 1GHz)

Podziękowania

Autor pragnie podziękować pani prof. dr hab. M. Steffen-Batogowej za inspiracje i udostępnienie danych do testowania programów.

Literatura

Breiman L., Friedman J. H., Olshen R. A., Stone C. J. 1984. *Classification and regression trees*. Belmont, CA: Wadsworth Statistics/Probability Series.

Cichosz, P. 2000. *Systemy uczące się*. Warszawa: WNT.

Dutoit, T. 1997. *An introduction to text-to-speech synthesis*. Dordrecht/Boston/London: Kluwer Academic Publishers.

Ferguson, G. A., Takane, Y. 1997. *Analiza statystyczna w psychologii i pedagogice.* Warszawa: PWN.

Lapis, W. 2001. Wyznaczanie stopnia redukcji własności do dystynktywnych, w: Górny M., Nowak P. (red.) *Miscellanea informatologica*, Poznań: UAM, red., str. 141 – 158.

Roman, S. 1999. *Access. Baza danych – projektowanie i programowanie*. Helion, Gliwice.

Salford Systems 2003. http://www.salford-systems.com/whitepaper.html.

Wang, M., Q., Hirschberg, J. 1992. Automatic classification of intonational phrase boundaries. *Computer Speech and Language*, 6, str. 175 – 196.