Adam Pelikant
Wyższa Szkoła Informatyki,
Katedra Inżynierskich Zastosowań Informatyki

METODY REPREZENTACJI ATRYBUTÓW W ZADANIACH ZGŁĘBIANIA DANYCH

Streszczenie – Artykuł zawiera dyskusję metod związanych z reprezentacją atrybutów w procesie budowania aplikacji zgłębiania danych (Data Mining). Przedstawiono zagadnienie redukcji ilości wymiarów analizowanej przestrzeni zadania – liczby istotnych atrybutów, a także standaryzowania wymiarów do uogólnionego zakresu zmienności. Główną częścią pracy jest omówienie problemów związanych ze sposobami reprezentacji atrybutów. Związane jest to głównie z koniecznością dyskretyzacji danych ciągłych w modelach, które są dedykowane dla danych nieciągłych (dyskretnych) oraz ciągłej reprezentacji danych dyskretnych w modelach wymagających tego typu atrybutów. Przedstawione zostały konkurencyjne algorytmy począwszy od „naiwnych" przez bardziej rozbudowane dyskretyzacji wstępującej, zstępującej oraz opartej o dyskryminator Fishera. Wskazano na miejsce zastosowanie metod oceny z zastosowaniem kryterium separowalności lub krzywej ROC. W oparciu o przykłady wskazano cechy prezentowanych rozwiązań.

1 Wprowadzenie

Szybki rozwój oraz powszechna dostępność zarówno sprzętu, jak i oprogramowania prowadzą do bardzo szybkiego wzrostu liczby gromadzonych danych. Bazy o wymiarach liczonych w dziesiątkach czy setkach GB nie są niczym niezwykłym. W praktyce komercyjnej spotykamy rozwiązania o rozmiarach liczonych dziesiątkami TB. Ponieważ procesy gromadzenia danych są procesami ciągłymi i w zasadzie nigdy nie rezygnujemy z przechowywania danych historycznych, proces wzrostu objętości danych będzie narastał lawinowo. Te rozważania dotyczą tylko schematów relacyjnych, a nie dotyczą jeszcze szybciej rozwijających się źródeł związanych z siecią WEB, gdzie szacunki objętości zgromadzonych danych liczone w dziesiątkach czy setkach PB nie są wcale przesadzone.

Jakość przetwarzania w środowiskach transakcyjnych jest „mierzona" szybkością przetwarzania i pewnością przechowywania. Dane zawarte w tak dużych strukturach są bardzo trudne do oceny i wymagają do przeprowadzenia analizy specjalistycznych narzędzi. Ponadto musimy

zgodzić się z opinią, że dane nie są tożsame z informacją, ale mogą stanowić jej źródło, jeśli ją z nich tylko umiemy wydobyć. Stąd konstruowane są różnego rodzaju algorytmy zgłębiania danych, podejmowania decyzji. Wiele z nich wymaga atrybutów o charakterze ciągłym (grupowanie) inne preferują dane dyskretne (drzewa decyzyjne). Nie wszystkie atrybuty są równie istotne w procesie wydobywania wiedzy. Przyczyny tego faktu mogą być wielorakie. Podstawową, wydawałoby się trywialną jest fakt braku wpływu atrybutu na stawiane hipotezy. Kolejną może być brak zaufania, co do poprawności, prawdziwości wszystkich lub części wartości atrybutu. Stąd waga właściwego wstępnego przygotowania danych jest nie do przecenienia, ale niestety proces ten jest również bardzo trudny. Opracowanie algorytmów automatycznego wykonania oceny jakości jest z reguły procesem złożonym i wymaga od programisty posiadania rozległej wiedzy teoretycznej, ale również dużego doświadczenia praktycznego.

2 Standaryzacja

W przypadku, gdy analizowana przestrzeń cech zawiera atrybut, którego zakres wartości znacznie odbiega od pozostałych, wynik w dużej mierze może zależeć tylko i wyłącznie od tej zmiennej. Zdarzenia takie często mają miejsce podczas operacji na danych wyrażonych w różnych jednostkach. Przykładem może być grupowanie obiektów wyrażonych za pomocą dwóch wymiarów: szerokości wyrażonej w centymetrach [cm] oraz wysokości wyrażonej w metrach [m]. Algorytm operujący na takich danych będzie błędnie dokonywał klasyfikacji, zaniżając znaczenie zmiennej wyrażonej w jednostkach wyższego rzędu. Standaryzacja polega na transformacji cech analizowanej przestrzeni tak, aby uzyskać podobnie szerokie przedziały przyjmowanych wartości.

Standaryzacja klasyczna opiera się na doprowadzeniu dowolnego rozkładu o parametrach μ i σ do rozkładu, zwanego odtąd rozkładem standaryzowanym, o wartości oczekiwanej μ=0 i odchyleniu standardowym σ=1. Zmienną x zastępujemy zmienną standaryzowaną u, która ma rozkład N(0, 1), zgodnie z zależnością (1)

$$u = \frac{x - \mu}{\sigma} \tag{1}$$

Standaryzacja medianowa jest standaryzacją analogiczną do klasycznej, lecz wykorzystującą medianę oraz odchylenie medianowe. Medianą cechy X przyjmująca wartości $x_1, \ldots, x_n$, nazywamy środkowy

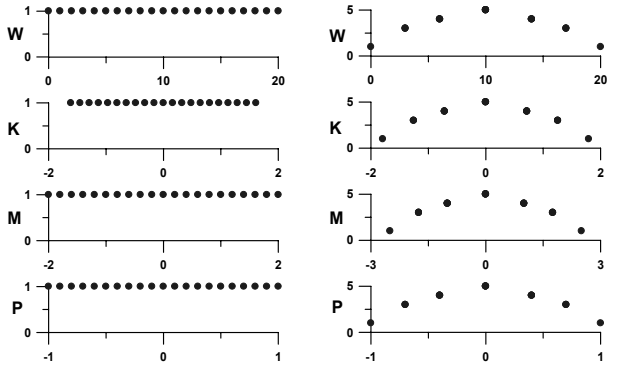
wyraz ciągu, gdy n jest liczbą nieparzystą, lub średnią arytmetyczną dwóch wyrazów środkowych, gdy n jest liczba parzystą (2).

$$m = \begin{cases} x_{k+1} & n = 2k+1 \\ \dfrac{x_k + x_{k+1}}{2} & n = 2k \end{cases} \tag{2}$$

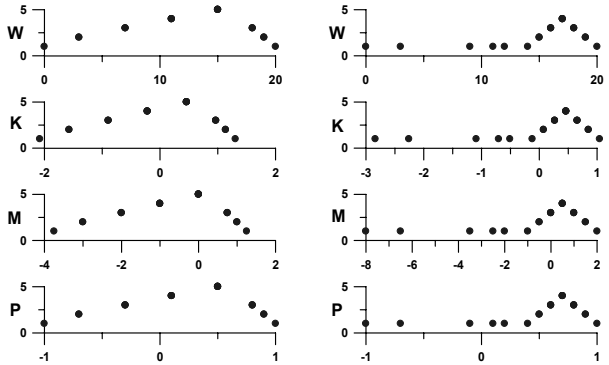
Standaryzacja skalująca liniowo do przedziału <0,1> jest szczególnym przypadkiem metody skalującej do przedziału <*min, max*>. Metoda ta przeprowadza transformację liniową danych pierwotnych do przedziału według wzoru (3):

$$u = \frac{x - min}{max - min}(max' - min') + min' \tag{3}$$

gdzie:
min, max - to wartość minimalna i maksymalna przedziału, w którym mieszczą się dane wejściowe
min', max' - wartość minimalna i maksymalna nowego przedziału, w którym mieścić się będą dane. Przykłady zastosowania przedstawionych algorytmów standaryzacji dla różnych charakterów wyjściowej dystrybucji częstotliwości występowania wartości cechy przedstawiają rysunki 1 i 2



Rys. 1. Skutek zastosowania do standaryzacji rozkładu wyjściowego W metody klasycznej K, medianowej M oraz przedziałowej P w przypadku rozkładu równomiernego A oraz normalnego B



Rys. 2. Skutek zastosowania do standaryzacji rozkładu wyjściowego W metody klasycznej K, medianowej M oraz przedziałowej P w przypadku rozkładu o dodatnim momencie rzędu 3 A oraz dużej zmienności gęstości danych B

Metoda przedziałowa zachowuje prawidłowe relacje pomiędzy wartościami, nie wprowadza także żadnych potencjalnych odchyleń od wartości. W pozostałych dwóch standaryzacjach rozkład danych wejściowych ma wpływ na uzyskiwane wyniki. Szczególnie jest to widoczne wtedy, kiedy poza obszarem o dużej koncentracji danych istnieją dane bardzo od nich odległe. W takim przypadku należy rozważyć możliwość usunięcia atrybutów oddalonych od głównego miejsca koncentracji, stosując kryterium 2σ lub 3σ. Pozostawiając tylko te, które mieszczą się w otoczeniu o promieniu dwóch (trzech) odchyleń standardowych od wartości średniej. Jednak przyjęcie takiego sposobu postępowania, jako ogólnej metody nie zawsze musi być poprawne (rozkłady złożone z wielu rozkładów normalnych). W takiej sytuacji musimy się zdać na doświadczenie, a czasem na intuicję.

3 Redukcja liczby arybutów

Pierwszym krokiem ustalenia, które z atrybutów są istotne z punktu widzenia analizy, może być wyznaczenie korelacji między każdymi parami argumentów. Dla atrybutów ciągłych możemy użyć współczynnika korelacji liniowej (4)

$$r_{jl} = \frac{\sum_{i=1}^{n} \left(x_{ij} - \bar{x}_j\right)\left(x_{il} - \bar{x}_l\right)}{nS(x_j) \cdot S\left(x_l\right)} \quad \left(j,l = 1,...,m\right) \tag{4}$$

natomiast dla atrybutów, cech dyskretnych (kategorycznych) współczynnika Spearmana (5)

$$R_{jl} = 1 - \frac{6\sum d_i^2}{n^3 - n} \tag{5}$$

gdzie:
d_i - odległość $(j, l = 1,...,m)$ między rangami cech (z reguły stosowany jest tzw. ranking gęsty uwzględniający pozycje ex aequo) "j" i "l" w *i*-tym obiekcie.
Możliwe jest stosowanie współczynnika zbiorowości Czuprowa (6),

$$T_{jl} = \sqrt{\frac{\chi^2}{m\sqrt{(m' - 1)(m'' - 1)}}} \quad (j, l = 1,...,m) \tag{6}$$

w którym test χ^2 dopasowania prawdopodobieństw hipotez do ich częstości względnych, jest wyznaczany według (7):

$$\chi^2 = \sum_{l'} \sum_{j'} \frac{\left(n_{j'l'} - \hat{n}_{j'l'}\right)^2}{\hat{n}_{j'l'}} \quad \left(n_{j'l'} - \hat{n}_{j'l'}\right) \tag{7}$$

gdzie: $n_{j'l'}$ - liczba obiektów (empiryczna) posiadających *j′*-tą odmianę cechy "j" oraz *l′*-tą odmianę cechy "l", przy czym w przypadku analizy podobieństwa cech możemy stosować miary Hellwiga (8)

$$d_{jl} = \sqrt{1 - |r_{jl}|} \tag{8}$$

lub miary opracowanej w szkole krakowskiej (T. Grabiński, S. Wydymus, A. Zeliaś) (9)

$$d_{jl} = \sqrt{1 - r_{jl}^2} \tag{9}$$

Wnioskowanie z takiej analizy jest dość ograniczone, pozwala tylko na eliminację tych argumentów, dla których korelacja ze wszystkimi innymi jest zero lub bliska zeru. Natomiast dla modułu korelacji równego 1 pozwala dwa argumenty traktować, jako tożsame. Natomiast miarą przyczynowo - skutkowości może być wielkość zysku informacji Gain (A, b) względem Gain (B, a) albo inna miara wsparcia (10)

$$supp = \frac{P(a \mid b)}{P(b \mid a)} \tag{10}$$

Każda z tych propozycji nie stanowi jednak pełnego rozwiązania problemu, wnosząc ponadto kolejne przetwarzania o bardzo wysokiej czasochłonności. Waga rozwiązania tego problemu jest bardzo duża – pozwala, bowiem na odkrywanie związków między danymi, z istnienia, których nie zdajemy sobie sprawy. Czyli otrzymujemy nową informację, szczególnie ważną dla procesów, zjawisk opisywanych dużą liczbą danych (zmiennych). Prowadzi to w istocie do pozyskania nowej wiedzy.

4 Zmiana typów atrybutów

Jeżeli rozważymy dystrybucje wartości atrybutu ciągłego w przestrzeni R^1, o liczbie wartości m w analizowanym przedziale, to koleje podziały na grupy, klasy możemy otrzymać dokonując rekursywnego podziału w wartościach progowych. Zakładamy ponadto znajomość przynależności tych atrybutów do klas C_i i=1..n. Takie podejście określane mianem dyskretyzacji wstępującej otrzymujemy bazując na minimalizowaniu entropii dzielonego podobszaru na skutek wprowadzenia progu podziałowego s (11),

$$ES(a,s) = \frac{m_{a\leq s}}{m} E(a \leq s) + \frac{m_{a>s}}{m} E(a > s) \tag{11}$$

w której E(a≤s) jest opisywana przez (12)

$$E(a \leq s) = \sum_{c_i} \left(-\frac{m_{a\leq s}^{C_i}}{m_{a\leq s}} \log_2 \frac{m_{a\leq s}^{C_i}}{m_{a\leq s}} \right) \tag{12}$$

gdzie: m – liczba atrybutów w analizowanym przedziale, $m_{a\leq s}$ – liczba atrybutów poniżej progu s, $m_{a\leq s}^{C_i}$ – liczba atrybutów poniżej progu s należących do klasy C_i. Wartość wyrażenia E(a>s) jest wyznaczana przez odpowiednią zmianę nierówności.
Kryterium zatrzymania procesu może być albo osiągnięcie górnej liczby podziałów zadeklarowanej przez użytkownika albo minimalnego progu zysku informacji na skutek wprowadzenia podziału s.

$$\Delta E = E(a) - ES(a,s) \tag{13}$$

Entropia jest miarą średniej ilości informacji, która przypada na zajście określonego zdarzenia ze zbioru danych S. Ogólny wzór na wartość entropii w przypadku analizowania zagadnień o n wartościach hipotezy przyjmuje postać (14):

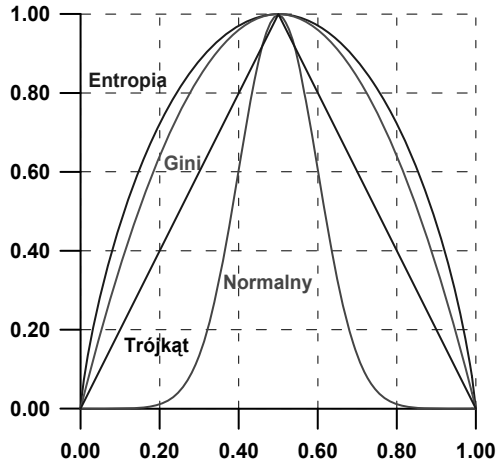
$$E(S) = -\sum_{i=1}^{n} p(i)\log_n p(i) \tag{14}$$

gdzie: p(i) – jest prawdopodobieństwem zajścia i-tego zdarzenia.
Oprócz standardowego wyznaczania entropii w oparciu o teorię Shanona (14) można również zastosować tzw. indeks Giniego (18).

$$E(p_1, p_2, ..., p_n) = p_1 \cdot p_2 \cdot ... \cdot p_n k_n \tag{15}$$

gdzie: S_v – zbiór obiektów w S z wartością atrybutu a = v.

Dozwolone jest dowolne kodowanie przekształcające przedział prawdopodobieństw na przedział wartości $\aleph$:<0,1>$\rightarrow$<0,1> aby dla prawdopodobieństw 0 i 1 funkcja miała wartość minimalną (zwykle 0) oraz posiadała jedno maksimum we wnętrzu przedziału, dzielące go na dwa podprzedziały, w których funkcja ta jest ściśle monotoniczna. W związku z tym, poza wymienionymi możliwe jest stosowanie np.: krzywej trójkątnej lub rozkładu normalnego (rys. 3)



Rys. 3. Przykładowe rozkłady wykorzystywane do obliczania zysku informacji

Kolejnym podejściem może być zastosowanie naiwnego klasyfikatora Bayesa do definiowania progów w dyskretyzacji zstępującej. Zakłada się, że jeżeli dana jest przestrzeń danych S i prawdopodobieństwa P zajścia hipotez $h_i \subseteq S$ i = 1,…,n $n \geq 1$ takich, że $\forall h_i \;\; P(h_i) > 0,$ to:

— hipotezy h_i są wzajemnie wykluczające się $\forall h_i \cap h_j = \varnothing$,
— hipotezy h_i są wspólnie wyczerpujące $\sum h_i = S_h$,

a atrybuty $e_{j_1},...,e_{j_k} \subseteq \Omega$ zwane symptomami dla $\{j_1,...,j_k\} \subseteq \{1,...,m\}$ $1 \le k \le m,$ oraz $m \ge 1$ takiego, że symptomy $e_{j_1},...,e_{j_k}$ są niezależne warunkowo, względem każdej hipotezy h_i, zachodzi zależność (17)

$$P(h_i \mid e_{j_1} \cap ... \cap e_{j_k}) = \frac{P(e_{j_1} \mid h_i) \cdot ... \cdot P(e_{j_k} \mid h_i) \cdot P(h_i)}{\sum_{l=1}^{n} P(e_{j_1} \mid h_l) \cdot ... \cdot P(e_{j_k} \mid h_l) \cdot P(h_l)} \tag{17}$$

W przypadku tego modelu konieczna jest znajomość tylko m*n prawdopodobieństw warunkowych $P(e_j|h_i)$ oraz n-1 prawdopodobieństw $P(h_i)$, które w prosty sposób mogą być obliczone z danych statystycznych. W wielu praktycznych podejściach, problem redukowany jest do dwóch hipotez, a przedział zmienności atrybutu, symptomu jest dzielony na dwie wartości e oraz $\neg$e.

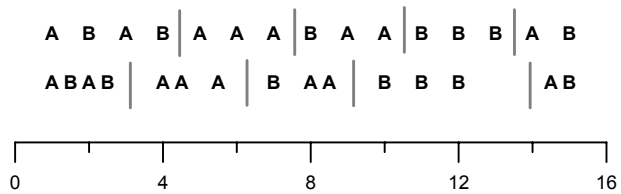
Konkurencyjnym podejściem jest rozpoczęcie procesu dyskretyzacji od podziału wszystkich atrybutów na jednoelementowe klasy, a następnie rekurencyjne ich łączenie w oparciu o test χ^2 wyznaczając jego wartość dla dwóch sąsiadujących przedziałów z_1 i z_2 po ich połączeniu (18)

$$\chi^2_{z_1,z_2} = \sum_{c_i} \frac{(m^{c_i}_{z_1} - e^{c_i}_{z_1})^2}{e^{c_i}_{z_1}} + \sum_{c_i} \frac{(m^{c_i}_{z_2} - e^{c_i}_{z_2})^2}{e^{c_i}_{z_2}} \tag{18}$$

W tym ujęciu wartość oczekiwane w obrębie łączonych przedziałów możemy oszacować, jako (19)

$$e^{c_i}_{z_2} = m_{z_2} \frac{m^{c_i}_{z_1 \cup z_2}}{m_{z_1 \cup z_2}} \qquad e^{c_i}_{z_1} = m_{z_1} \frac{m^{c_i}_{z_1 \cup z_2}}{m_{z_1 \cup z_2}} \tag{19}$$

Łączenie przedziałów zostanie zakończone, kiedy poziom testu χ^2 przy próbie łączenia dowolnych przedziałów przekroczy zadany próg lub osiągniemy zadaną minimalną liczbę klas (np. przyłączenie wszystkich klas mniej licznych niż zadana ilość). Obie metody dyskretyzacji zakładają istnienie hipotezy warunkującej przypisanie atrybutu do klasy C_i, której wartość jest wykorzystywana do oceny dyskretyzacji. Obie natomiast abstrahują od odległości między poszczególnymi atrybutami analizując tylko ich gęstość, sąsiedztwo rys. 4.



Rys. 4. Przykładowe warianty dyskretyzacji: tylko z uwzględnieniem gęstości oraz wraz z uwzględnieniem położenia

Innym wariantem może być zastosowanie metod nienadzorowanych takich jak grupowanie. Zastosowanie metod typu K-means lub C-means z narzuconą liczbą klas może okazać się niezadowalające stąd możliwe zastosowanie metody grupowania w oparciu o funkcję górską. Metoda ta zakłada tylko wstępny podział analizowanego przedziału, dla którego to wyznaczamy podstawową postać funkcji górskiej *M*, dla każdego punktu N_{ij}, (X_i, Y_j) w zbiorze *N* o postaci (20)

$$M(N_{ij}) = \sum_{k=1}^{q} e^{-(\alpha d(N_{ij}, O_k))} \tag{20}$$

przy czym O_k jest *k*-tym punktem danych (x_k, y_k), □ jest stałą dodatnią i $d(N_{ij}, O_k)$ jest miarą odległości miedzy N_{ij} i *Ok*. Dla nowo powstałych węzłów skorygowana funkcja górska ma postać (21)

$$M_{k+1}(N_{ij}) = M_k(N_{ij}) - M_k^* e^{-\beta d(N_2^*, N_{ij})} \tag{21}$$

Wynikiem jest liczba środków wraz z funkcjami przynależności, które można przekształcić w „ostre" klasy odcinając na poziomie z.

Kolejną propozycją dyskretyzacji może być zastosowanie dyskryminacji liniowej, która polega na wyznaczeniu dwóch hiperpłaszczyzn określających margines między dwoma klasami. Przynależność, do nich jest kodowana, jako y_A=1 oraz y_B=-1. Podział taki można opisać równaniem (22)

$$\mathbf{w}^T \mathbf{x} + \mathbf{b} = \pm 1 \tag{22}$$

Uzyskanie najlepszej separowalności klas uzyskujemy dzięki maksymalizacji marginesu m (23)

$$m = \frac{2}{\|\mathbf{w}\|^2} \tag{23}$$

W praktyce prowadzi to do minimalizowania funkcji odwrotnej przez wyznaczenie miejsc zerowych Langranianu o postaci (24)

$$\mathcal{L} = \frac{1}{2}\mathbf{w}^T\mathbf{w} + \sum_{i=1}^{n}\alpha_i\left(1 - y_i\left(\mathbf{w}^T x_i + b\right)\right) \tag{24}$$

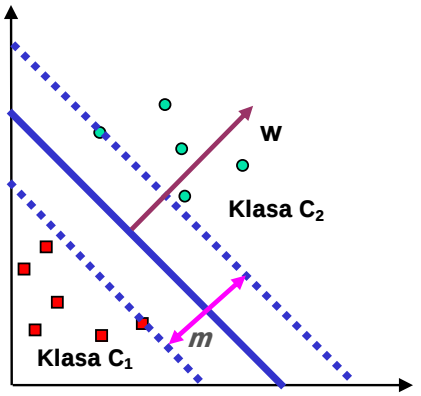
Pociąga to za sobą konieczność rozwiązania zadania programowania nieliniowego (25)

$$\max\ W(\alpha) = \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{j=1}^{n}\sum_{i=1}^{n}\alpha_i\alpha_j y_i y_j x_i x_j \tag{25}$$

z ograniczeniami (26)

$$\alpha_i \geq 0, \quad \sum_{i=1}^{n}\alpha_i y_i = 0 \tag{26}$$

Przykładowy podział zbioru danych uzyskany dzięki zastosowaniu metody liniowej dyskryminacji Fishera przedstawia rys. 5.



Rys. 5. Ilustracja zastosowania liniowej klasyfikacji z zastosowaniem SVM

Takie ujęcie dla zadania jednowymiarowego daje dobre rezultaty, kiedy atrybuty dwóch klas są uszeregowane – monotoniczne. Daje to rozwiązanie trywialne, do którego nie ma sensu używać tak złożonego aparatu matematycznego. Jednak wprowadzenie tzw. funkcji jądra powodujących przeniesienie analizy zadania nad R^1 do przestrzeni o wyższym wymiarze daje dużo lepsze rezultaty. Do najczęściej stosowanych kerneli należą: wielomianowe stopnia d (27)

$$K(\mathbf{x},\mathbf{y}) = (\mathbf{x}^T\mathbf{y}+1)^d \tag{27}$$

radialne, oparte o odchylenie standardowe σ (28)

$$K(\mathbf{x},\mathbf{y}) = \exp(\frac{-\|\mathbf{x}-\mathbf{y}\|^2}{2\sigma^2}) \tag{28}$$

sigmoidalne (29)

$$K(\mathbf{x},\mathbf{y}) = \tanh(k\mathbf{x}^T\mathbf{y}+\theta) \tag{29}$$

gdzie k, θ wybrane stałe przekształcenia (nie dla wszystkich ich wartości gwarantowana jest zbieżność rozwiązania)
Poważnym problemem jest obiektywna ocena jakości dyskretyzacji, która opierałaby się bezpośrednio na położeniu danych i grup w przestrzeni. Taki problem nie istnieje tylko w przypadku liniowego dyskryminatora Fishera, w który oceną miary jest szerokość marginesu. W pozostałych przypadkach oceny takiej możemy dokonać na podstawie

- ilorazu średniej odległości elementów w grupie i średniej odległości grup (30)

$$e_1 = \frac{\frac{1}{n}\sum_{k=1}^{L} mean(d(v_i,v_j))}{mean(d(v_p,v_q))} \tag{30}$$

- ilorazu sumy "momentów bezwładności" grup i "momentu bezwładności" wszystkich elementów (31)

$$e_2 = \frac{\sum_{k=1}^{L} I_k}{I} \tag{31}$$

- funkcji oceny grupowania wykorzystującej sumę miar odległości przykładów od grup, do których te przykłady zostały zaklasyfikowane (32)

$$e(C) = \sum_{c\in C}\sum_{x\in X} f(x,c)\cdot u(x,c) \tag{32}$$

gdzie *f(x,c)* jest wybraną funkcją odległości *u(x,c)* i określa stopień przynależności przykładu *x* do grupy *c*. Dla grup twardych $u(x,c) \in \{0,1\}$ dla rozmytych $u(x,c) \in [0,1]$.
Wadą funkcji postaci (32) dla wielu funkcji przynależności jest brak jej monotoniczności (33)

$$e(C_1) < e(C_2) \text{ gdy } |C_1| > |C_2| \tag{33}$$

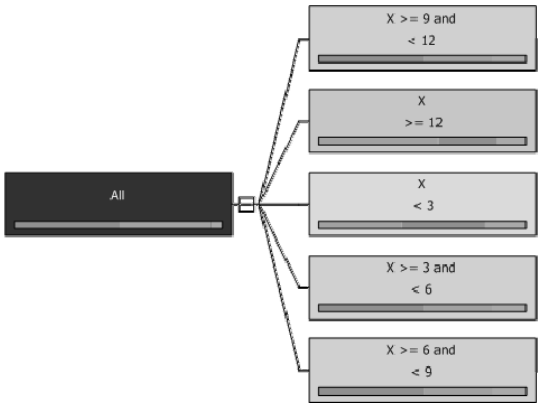
gdzie *C*1, *C*2 są dwoma proponowanymi przez algorytm grupami zbioru danych.
W krańcowym przypadku łatwo jest zauważyć, że grupowanie, w którym każdy przykład jest w odrębnej, jednoelementowej grupie, jest oceniane najlepiej. Dlatego w algorytmach, które nie mają ustalonej a priori liczby grup, czynnik reprezentujący liczbę elementów w grupie trzeba włączyć w funkcje oceny. Jedną z takich funkcji uwzględniających liczbę klastrów jest Bayesowskie Kryterium Informacji (*Bayesian Information Criterion*, BIC) (34)

$$BIC(C) = -2 \cdot e(C) + v(C) \cdot \log N \tag{34}$$

gdzie *v(C)* jest liczbą parametrów modelu *C*, a N to wielkość zbioru danych.
Funkcja ta porównuje zysk informacji, jaki uzyskujemy ze zwiększonej liczby grup, w porównaniu z wiążącym się z tym wzrostem skomplikowania opisu modelu.

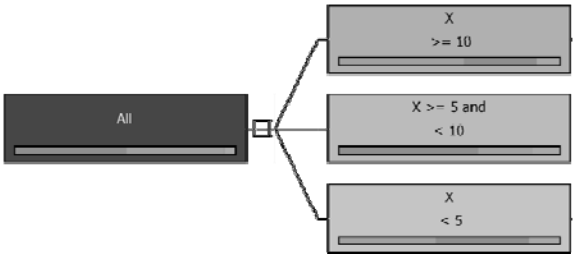
5 Testy numeryczne

Już proste testy dla danych jednowymiarowych przedstawionych na rysunku 4 wskazują na wagę podejmowanego problemu. W rozwiązaniach komercyjnych królują rozwiązania wymagające statycznej definicji liczby podziałów. Nawet wtedy, kiedy używana jest metoda oparta o grupowanie, ze względu na stosowanie algorytmu K-means liczba grup dawana jest statycznie (rys. 6). Prowadzi to do konieczności wielokrotnej walidacji otrzymanego drzewa. Mimo wszystko taki podział generuje słabe drzewa, co jest widoczne na histogramach umieszczonych na paskach w węzłach drzewa.



Rys. 6. Drzewo decyzyjne z naiwnym wariantem dyskretyzacji dla ustalonej liczby klastrów równej 5

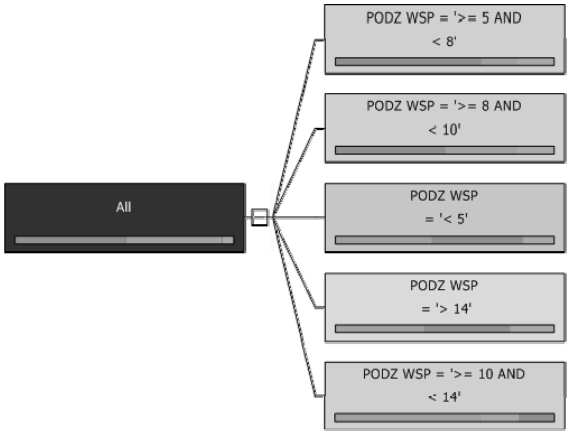
Przyjęcie mniejszej liczby podziałów zdecydowanie poprawia dystrybucję danych w węzłach (liściach) drzewa (rys.7). Niestety bardzo irytujące jest przyjęcie jedno elementowej reprezentacji wartości NULL, nawet wtedy, gdy nie występuje w źródle.



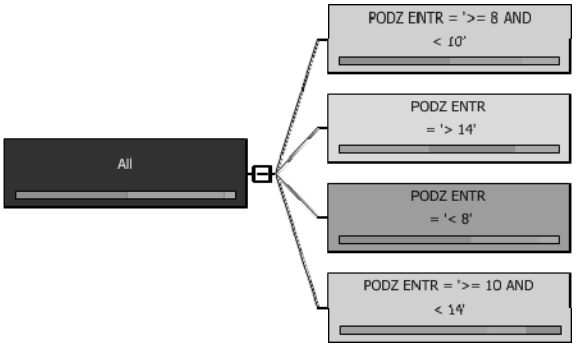
Rys. 7. Drzewo decyzyjne z naiwnym wariantem dyskretyzacji dla ustalonej liczby klastrów równej 3.

Zastosowanie algorytmów zstępującego i wstępującego do przykładowych danych jednowymiarowych daje jeden stan taki, że dyskretyzacje otrzymane obydwiema metodami dają taki sam rezultat. Wygenerowane dla nich drzewo decyzyjne przedstawia rys. 8. Widoczne jest przez histogramy słabe dopasowanie drzewa, ale i tak lepsze niż dla wbudowanej metody dyskretyzacji.

Dla kolejnego kroku podziału (łączenia) oba algorytmy, co oczywiste, generują odmienne podziały. Jednak w obu przypadkach dystrybucja klas w obrębie histogramów wskazuje dużo lepsze dopasowanie do danych. Na rysunku 9 pokazano drzewo wynikające z kolejnego podziału dla metody zstępującej, natomiast rys 10. przedstawia kolejny krok łączenia dla metody wstępującej



Rys. 8. Drzewo decyzyjne na wspólnym poziomie dyskretyzacji uzyskanym metodą zstępującą i wstępującą

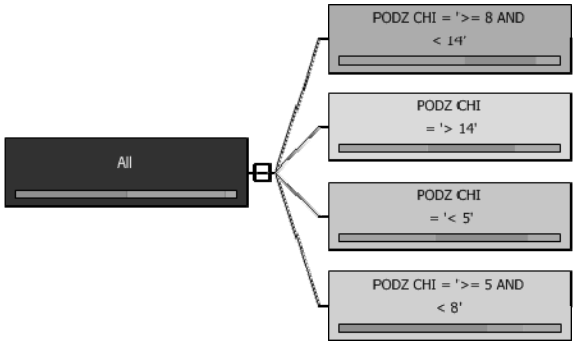


Rys. 9. Drzewo decyzyjne dla dyskretyzacji metodą wstępującą (entropia) krok przed osiągnięciem poziomu wspólnego

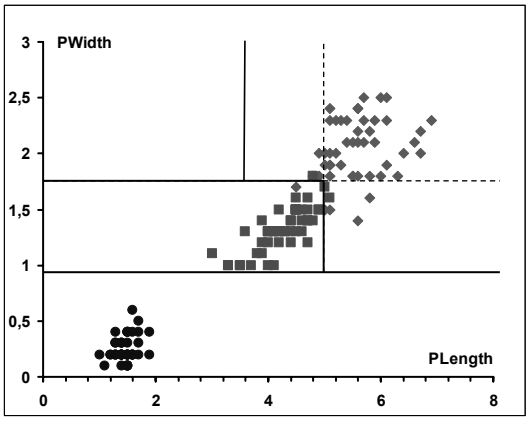
Wybór pomiędzy wynikami uzyskanymi różnymi metodami musi zostać dokonany arbitralnie, ponieważ nie istnieje wspólna miara jakości rozwiązania.

Dane testowe zostały z premedytacją dobrane tak, aby dawały słabą dyskryminację. Pozwala to wnioskować, że dla każdego lepiej uwarunkowanego zadania podziały będą generowane lepiej.

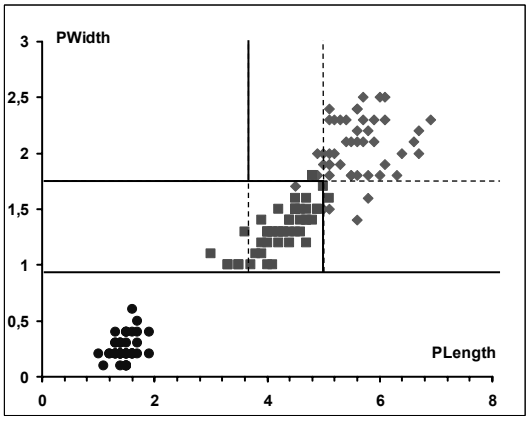
Podobne rezultaty można uzyskać analizując skutki zastosowania metod dyskretyzacji dla dobrze znanych danych testowych, treningowych pochodzących z repozytorium zgłębiania danych – Iris.



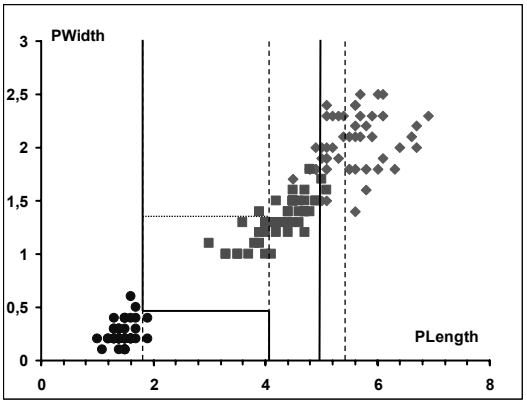
Rys. 10. Drzewo decyzyjne dla dyskretyzacji metodą zstępującą krok po osiągnięciu poziomu wspólnego



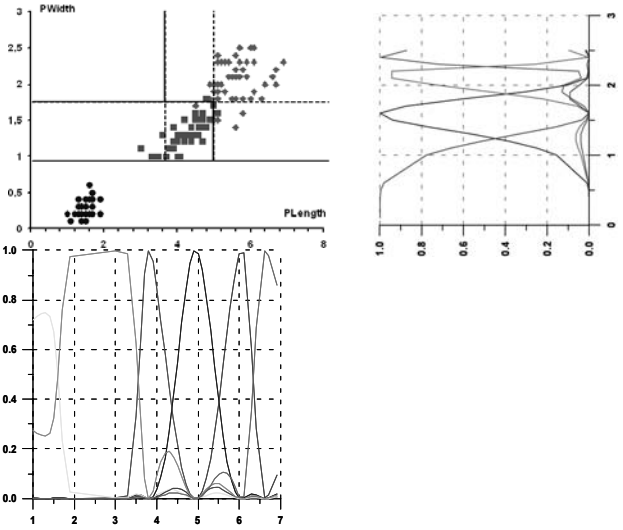
Rys. 11. Ilustracja podziału zbioru Iris za pomocą współczynnika Ginni dla trzech grup



Rys. 12. Ilustracja podziału zbioru Iris za pomocą entropii dla trzech grup



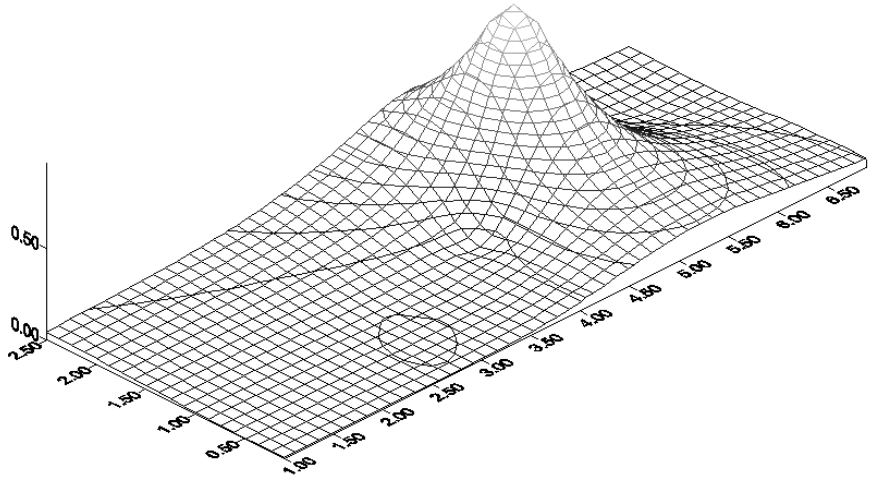
Rys. 13. Skutek przyjęcia złej liczby podziałów dla klasyfikatora Bayesa



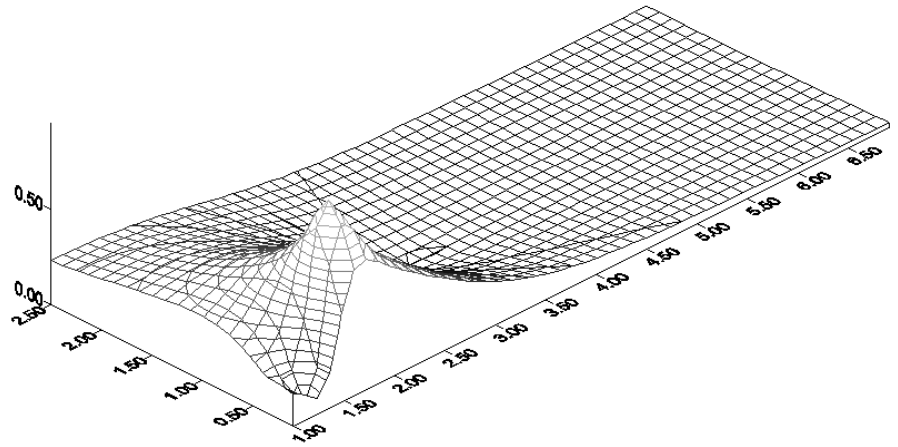
Rys. 14. Ilustracja podziału atrybutów Plength i Pwidth zbioru Iris za pomocą złożenia jednowymiarowego grupowania algorytmem górskim

Zastosowana została dyskretyzacja wstępująca, wykorzystująca wskaźnik Ginni oraz entropię. W obu przypadkach otrzymano w zasadzie taką samą końcową klasyfikację (rys. 11 i 12). Różnice wynikają jedynie z kolejności wykonywania podziałów – linie przerywane. W celu pokazania znaczenia poprawnej dyskretyzacji pokazano klasyfikację opartą o źle dobraną liczbę podziałów – 4 klastry (rys. 13). Poprawy jakości podziału możemy dokonać stosując dla każdego atrybutu (każdej osi na wykresie) jednowymiarowego

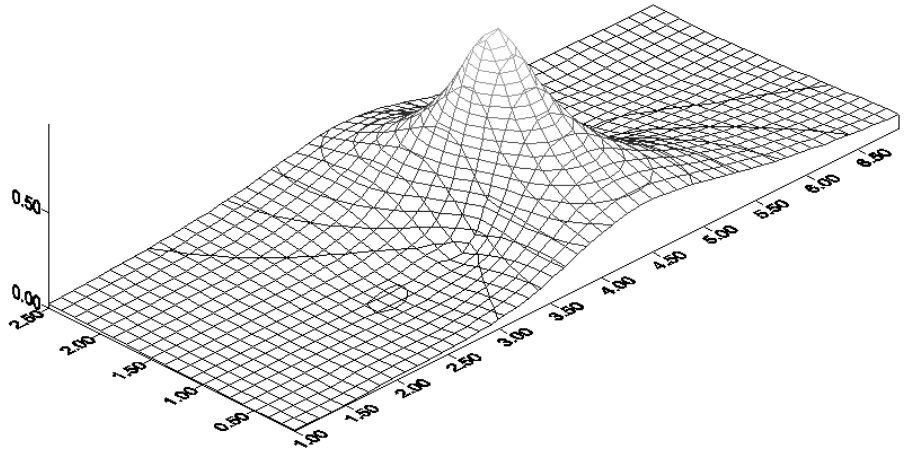
grupowania z zastosowaniem algorytmu górskiego (rys. 14). W takim przypadku złożenie klastrów jednowymiarowych do klastrów w przestrzeni R^2 prowadzi do utworzenia dobrze uwarunkowanego drzewa.



Rys. 15. Ilustracja podziału atrybutów Plength i Pwidth zbioru Iris za pomocą dwuwymiarowego grupowania algorytmem górskim (jeden z klastrów)



Rys. 16. Ilustracja podziału atrybutów Plength i Pwidth zbioru Iris za pomocą dwuwymiarowego grupowania algorytmem górskim (jeden z klastrów)



Rys. 17. Ilustracja podziału atrybutów Plength i Pwidth zbioru Iris za pomocą dwuwymiarowego grupowania algorytmem górskim (jeden z klastrów)

Możliwe jest również rozwiązanie zagadnienia grupowania bezpośrednio w przestrzeni dwuwymiarowej. Czyli nie dokonujemy dyskretyzacji atrybutów opierając się na ich dystrybucji względem każdego z parametrów, ale analizując je łącznie. Wyniki grupowania dwuwymiarowego dla testowego zbioru Iris przedstawiają rysunki 15 – 16. Takie podejście do wstępnego przetworzenia powoduje praktycznie natychmiastowe uzyskanie liści drzewa.

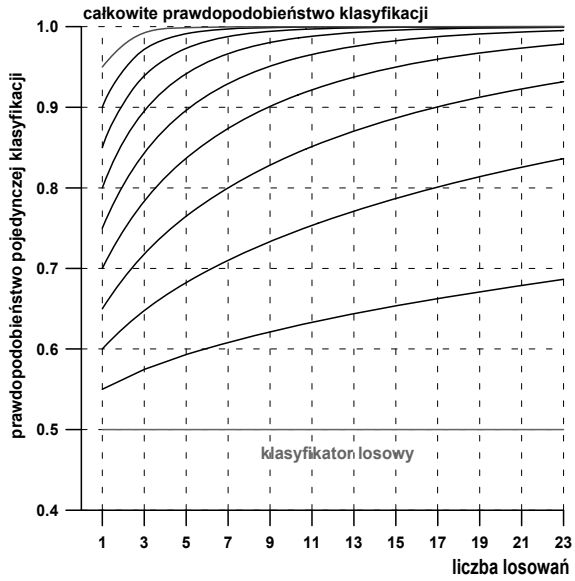
Niestety nie możemy spodziewać się równie pozytywnych rezultatów w przypadku ogólnym. Bardzo dobre wyniki uzyskane dla zbioru Iris wynikają z dobrej „podatności" tych danych na stosowanie metod uczenia bez nauczyciela, do których należy grupowanie.

6 Podsumowanie

Metody reprezentacji danych w algorytmach decyzyjnych mają decydujący wpływ na sposób tworzenia reguł. Pociąga to za sobą konieczność weryfikacji przyjętych rozwiązań na podstawie zbiorów testujących odpowiednio wyekstrahowanych z danych źródłowych. Innym podejściem może być stosowanie algorytmów wzmacniających uczenie – bagging, boosting czy adaboosting. W najprostszym przypadku zastosowania algorytmu boosting (losowania ze zwracaniem) wzmocnienie poziomu prawdopodobieństwa poprawnej klasyfikacji zmienia się według zależności (35).

$$p_c = \sum_{i=k+1}^{2k+1} \binom{2k+1}{i} p^i (1-p)^{2k+1-i} \tag{35}$$

Gdzie p jest prawdopodobieństwem poprawnej klasyfikacji dla pojedynczego klasyfikatora. Należy zauważyć, że prawdopodobieństwo klasyfikacji losowej ma wartość 0.5. Zatem zastosowanie nawet słabych klasyfikatorów, przy stosunkowo niewielkiej ich liczbie, daje szybko duże wzmocnienie (rys. 18). Dla 10 klasyfikatorów przy pojedynczej szansie poprawnej klasyfikacji 0.8, prawdopodobieństwo dla lasu klasyfikatorów wyniesie już ok. 0.99.

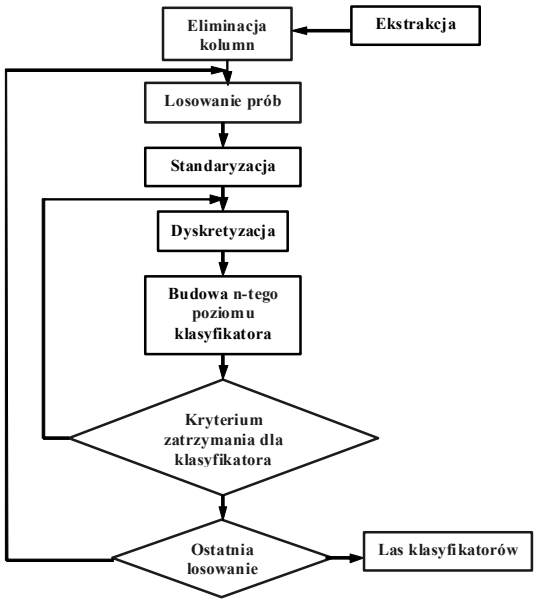


Rys. 18. Wzrost prawdopodobieństwa poprawnej klasyfikacji w metodzie boosting przy 2*k+1 losowaniach i stałej wartości poprawnej klasyfikacji pojedynczego klasyfikatora

Metody te mogą być stosowane nie tylko do weryfikowania końcowych reguł, ale również mogą być stosowane do poprawy, jakości metod dyskretyzacji.

Na podstawie przedstawionych rozważań ogólnych możemy spróbować pokusić się o zbudowanie schematu blokowego uogólnionego algorytmu zgłębiania danych (rys. 19). Na uwagę zasługuje cykliczna dyskretyzacja atrybutów na każdym poziomie budowy klasyfikatora. W przypadku drzewa decyzyjnego spowoduje to, że w każdym węźle, podział względem atrybutu może wystąpić dla różnych jego wartości. Również cyklicznie odbywa się losowanie próbki

danych testowych oraz budowa każdego z elementarnych klasyfikatorów lasu.



Rys. 19. Postulowany algorytm zgłębiania z zastosowaniem lasu klasyfikatorów

W obecnej chwili, pomimo opracowania algorytmów dla każdego z elementarnych bloków, elementarnych zadań, nie zakończono jeszcze budowy całego systemu. Proces integracji elementów składowych w uogólnioną metodę nie jest trywialny i wymaga rozwiązania szeregu problemów zasygnalizowanych w tym artykule. Wiele z tych zagadnień wymaga podejścia kontekstowego. Kontekstowość wyboru kryteriów lub algorytmów, wymaga bardzo rozległych podstawowych badań teoretycznych.

Literatura

[1] Kowalczyk A., Pelikant A.:: *Fuzzy clustering in relational databases*, XII International Conference - System Modelling and Control SMC'2007

[2] Pelikant: A.: *Bazy danych w zastosowaniach praktycznych*. roz. Kierunki rozwoju baz danych i technologii z nimi związanych, monografia WSInf

[3] Agata M., Pelikant: A.: *Support methods for weak learning algorithms – Adaboost*, XII International Conference - System Modelling and Control SMC'2007

[4] Gomide F., Silva L., Yager R.: *Participatory Learning Clustering, Forecasting and Evolutionary Fuzzy Systems*, BISCSE'05 University of California, Berkeley, November 2005

[5] Ahmad A., Dey L.: *A k-mean clustering algorithm for mixed numeric and categorical data*, Data & Knowledge Engineering 63 (2007) 503–527 Elsevier

[6] Vapnik V.: T*he Nature of Statistical Learning Theory*. Springer-Verlag, 1995.

[7] Vapnik V., Kotz S.: *Estimation of Dependences Based on Empirical Data*, Springer, 2006.

[8] Grąbczewski K., Duch W.: *The separability of split value criterion*. Proceedings of the 5th Conference on Neural Networks and Their Applications, s 201–208, Zakopane, Poland, 2000.

[9] Grąbczewski K., Duch W.: *Heterogeneous forests of decision trees*. Proceedings of International Conference on Artificial Neural Networks, Vol.2415 seria Lecture Notes in Computer Science, ss 504–509. Springer, 2002.

[10] Dietterich T. G.: *Approximate statistical tests for comparing supervised classification learning algorithms*. Neural Computation, 10(7):1895–1924, 1998.

[11] Bruha I., Berka P.: *Discretization and fuzzification of numerical attributes in attribute-based learning*. Fuzzy Systems in Medicine, wolumen 41 serii Studies in Fuzziness and Soft Computing, strony 112–138. Physica-Verlag (Springer), Heidelberg, 2000.

[12] Freund Y., Schapire R.: *Experiments with a new boosting algorithm. Machine Learning*: Proceedings of the Thirteenth International Conference, ss 148–156, 1996.

[13] Freund Y., Schapire R.: *A decision-theoretic generalization of on-line learning and an application to boosting*. Journal of Computer and System Sciences, 55(1):119–139, 1997.

[14] John G. H., Langley P.: *Estimating continuous distributions in Bayesian classifiers*. Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence, San Mateo, 1995. Morgan Kaufmann Publishers.

[15] Wilson D. R., Martinez T. R.: *Improved heterogeneous distance functions*, Journal of Artificial Intelligence Research, 11:1–34, 1997.

[16] Li B., Shen Y., Li B.: *A New Algorithm for Computing the Minimum Hausdorff Distance Between Two Point Sets on A Line Under Translation,* Information Processing Letters (2007).

METHODS OF ATRIBUTES REPRESENTATION IN DATA MINING PROBLEMS

Summary – The paper contains the discussion of methods connected with attributes representation in the process of build data mining applications. The task of dimensions number reduction in the analyzed problem space was introduced and presents – the number of essential attributes, and standardization to generalized range variation. The main part of this work is description of the problems with attributes representation. This is connected mainly with necessity of continuous data discreetisation in mining models which are dedicated for non - continuous data (discreet) as well as the continuous representation of discrete data in exacting this attribute type mining models. Competitive algorithms were introduced begin from simple „naive" by more extending discreetisation algorithms ascending, descending as well Fisher's discriminator. The place of methods use and evaluation due to separate criterion or curve the ROC was pointed. Based on examples the features of presented methods and algorithms were shown.