ZESZYTY NAUKOWE POLITECHNIKI ŚLĄSKIEJ 2002
Seria: GÓRNICTWO z. 255 Nr kol. 1564

Agnieszka KOWAL
Politechnika Śląska, Gliwice

ZMODYFIKOWANY ALGORYTM GRUPOWANIA PRÓBEK WĘGLA WOKÓŁ C-CENTROIDÓW

Streszczenie. W referacie przedstawiono zmodyfikowany algorytm grupowania próbek węgla wokół c-centroidów. Klasyfikacji poddawany jest zbiór 500 próbek węgla. Program tworzy żądaną liczbę klas oraz wyznacza rozmyte współczynniki przynależności próbek do klas. Optymalny podział zapewnia procedura autokorekty.

MODIFIED C-MEANS CLUSTERING ALGORITHM OF COAL SAMPLES

Summary. In the article modified c-means clustering algorithm of coal samples is presented. A file of 500 coal samples is divided. The program creates a required number of clusters and owns the fuzzy membership coefficients for the samples. A procedure of auto-correction assures optimal classification.

1. Wprowadzenie

Algorytm grupowania wokół c-centroidów jest przykładem analizy grupowania typu *clustering*. Analiza ta polega na poszukiwaniu grup zwanych ***klastrami*** wśród danych. Dąży do tworzenia grup, w których obiekty należące do tej samej grupy są maksymalnie podobne do siebie, podczas gdy obiekty w różnych grupach są maksymalnie różne.

Metody klasyfikacji typu clustering umożliwiają podział danych także w przypadku nieznajomości postaci klas na początku czy w przypadku dysponowania wielowymiarowym zbiorem danych.

Przeprowadzona analiza danych dąży do uzyskania *obiektywnej* i *stabilnej* klasyfikacji. Zadanie, które jest rozwiązywane w analizie ***typu clustering*** można sformułować następująco:

Dany jest zbiór n pojedynczych obiektów, z których każdy jest określony przez zbiór p zmiennych. Wyprowadź użyteczny podział na liczbę c klas. Zarówno liczba klas, jak i liczba zmiennych p opisujących obiekt jest określona (skończona).

2. Metody podziałów w ujęciu crisp

Metoda podziałów tworzy k klastrów poprzez klasyfikację danych w k grupach, które spełniają następujące wymagania podziału:

1. każda grupa musi zawierać przynajmniej jeden obiekt,
2. każdy obiekt musi należeć dokładnie do jednej grupy.

Warunki te określają:

1. istnienie co najwyżej tylu grup ile jest obiektów: $c \leq n$,
2. brak dzielenia obiektu między klastrami (różne klastry nie mogą mieć wspólnego obiektu),
3. pokrycie pełnego zbioru danych przez k klastrów.

Generalnie algorytmy znajdują „dobry" podział w sensie, że obiekty należące do jednego klastra powinny być podobne do siebie lub zależne względem siebie, natomiast obiekty różnych klastrów powinny być bardzo odmienne. Celem wówczas staje się odkrycie struktury już istniejącej wśród danych lub narzucenie nowej struktury lepiej spełniającej warunki podziału.

3. Metody rozmytego rozpoznawanie wzorców

Do grona klasycznych metod membership-roster (*ang.: fuzzy membership-roster method*) grupowania rozmytego opartych na rozmytych relacjach równoważności (*ang.: fuzzy equivalence relation-based hierarchical clustering method)* należy metoda grupowania wokół c-centroidów (*ang.: fuzzy c-means clustering*) zaproponowana przez Bezdeka w 1981 r. [1]. Algorytm rozmytego grupowania wokół c-centroidów polega na grupowaniu obiektów w oparciu o analizę odległości między obiektami w bazie. Jako kryterium podobieństwa badana jest suma kwadratów niepodobieństwa obiektów wewnątrz klastrów. Algorytm stosuje metodę graficznej odległości zwaną *single linkage algorithm* [Gower, 1969].

Każdą klasę definiuje standardowy wzorzec zwany ***prototypem***. Wzorzec poddany sklasyfikowaniu jest kolejno porównywany z określonymi prototypami klas. Stwierdza się stopień zgodności wzorca względem prototypów. Wzorzec jest przyłączany do kolejnych klas reprezentowanych przez prototypy w określonym stopniu odpowiadającym ich stopniowi zgodności. Efektywność tej metody zależy od reprezentatywności zbioru prototypów prezentujących klasy.

Niech n – liczba rozpoznanych i uporządkowanych w N_n klas wzorców. Dla danego wzorca wyznacza się: $\mu_k(x)$ - stopień zgodności x ze wzorcem reprezentującym *k-tą* klasę,

gdzie $x = [x_1, x_2, \ldots x_p]$ – badany wzorzec,

x_i - miara związana z *i-tą* właściwością wzorca.

Klasyczny algorytm rozmytego grupowania wokół c-centroidów wymaga określenia następujących parametrów: liczba klastrów c, gdzie $2 \le c \le n$; wykładnicza miara m, gdzie $1 < m < \infty$; symetryczna i dodatnio określona macierz G ($p \times p$) lub odległości pomiędzy każdą parą obiektów $d \in (1, \infty)$ i $d \in R$; metoda do utworzenia wartości początkowych macierzy $\tilde{U}^{(0)}$; kryterium końcowe $\Delta = \left\| \tilde{U}^{(l+1)} - \tilde{U}^{(l)} \right\|_G \le \varepsilon$.

Na ich podstawie jako rozwiązanie końcowe wybierany jest podział zbioru danych X na c klastry spełniający kryterium ε.

Algorytm obejmuje następujące kroki:

Krok 1: Wybierz liczbę klastrów c, taką, że $2 \le c \le n$ oraz wartość wagi m, taką, że $1 < m < \infty$ oraz symetryczną i dodatnio określoną macierz G ($p \times p$).
Obierz podział początkowy $\tilde{U}^{(0)} \in M_{fc}$, niech poziom $l = 0$.

Krok 2: Oblicz wartości c środków rozmytych klastrów $\{v_i^{(l)}\}$ stosując podział $\tilde{U}^{(l)}$ według zależności:

$$v_j = \frac{\sum_{k=1}^{n} \left[\mu_j(x_k)\right]^m x_k}{\sum_{k=1}^{n} \left[\mu_j(x_k)\right]^m}$$

Krok 3: Wyznacz nową macierz współczynników podziału $\tilde{U}^{(l+1)}$ wykorzystując wyliczone wartości $\{v_i^{(l)}\}$ jeżeli $x_k \neq v_i^{(l)}$, według zależności:

$$\mu_j(x_k) = \frac{\left(\frac{1}{\left\| x_k - v_j \right\|_G^2} \right)^{1/(m-1)}}{\sum_{j=1}^{c} \left(\frac{1}{\left\| x_k - v_j \right\|_G^2} \right)^{1/(m-1)}}, \qquad i = 1, \ldots, c; \quad k = 1, \ldots, n \tag{1}$$

Należy przyjąć dla każdego $x_k \in X$:

JEŻELI $||x_k - v_i(l)||2>0$ dla każdego $i \in Nc \Rightarrow$ $A_i^{(l+1)}(x_k) = \left[\sum_{j=1}^{c} \left(\frac{\left\| x_k - v_i^{(l)} \right\|^2}{\left\| x_k - v_j^{(l)} \right\|^2} \right)^{\frac{1}{m-1}} \right]^{-1}$

JEŻELI $||x_k - v_i^{(l)}||^2=0$ dla $i \in I \subseteq N_c \Rightarrow$ dla $i \in I$ $A_i^{(l+1)}(x_k)$ jest dowolną liczbą rzeczywistą nieujemną taką, że: $\sum_{i \in I} A_i^{(l+1)}(x_k) = 1$ dla $i \in N_c - I$ $A_i^{(l+1)}(x_k) = 0$.

Krok 4: Wybierz odpowiednią normę macierzy i wylicz $\Delta = \left\| \tilde{U}^{(l+1)} - \tilde{U}^{(l)} \right\|_G$, np. $\left| P^{(l+1)} - P^{(l)} \right| = \max_{i \in N_c, k \in N_n} \left| A_i^{(l+1)}(x_k) - A_i^{(l)}(x_k) \right|$

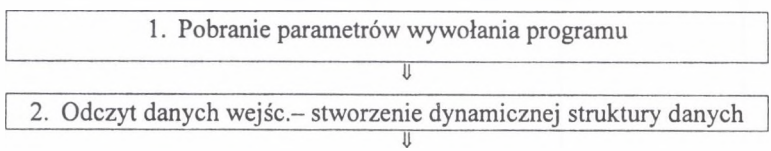
Jeżeli $\Delta > \varepsilon$, to zmień $l = l + 1$ i idź do kroku 2.

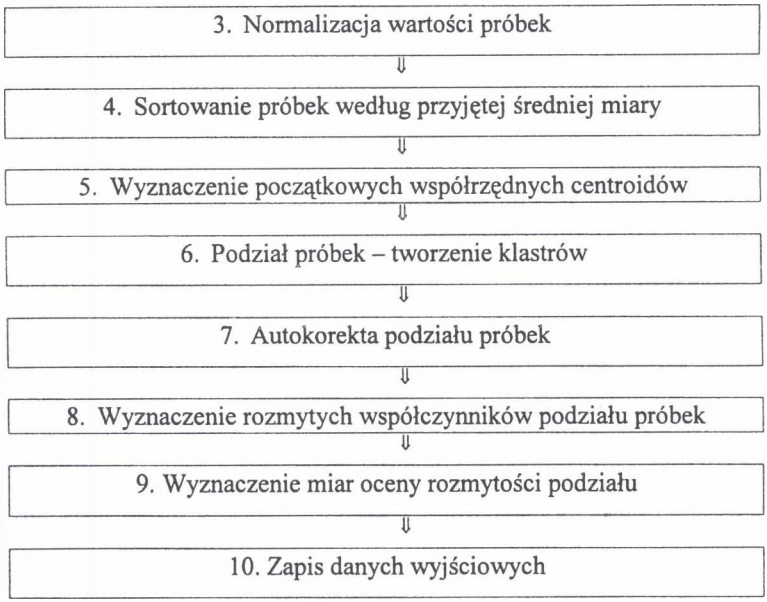
Jeżeli $\Delta \leq \varepsilon$, to koniec.

W algorytmie w pierwszym kroku przyjmuje się dowolny podział obiektów na c klastrów. Następnie wyznacza się c środków klastrów v_i dla przyjętej wagi m i uaktualnia podział według zasady: dla obiektu nie pokrywającego się ze środkiem klastra (odległość obiektu od klastra nie jest zerowa) oblicza się wartość współczynnika przynależności μ_{ik} według zależności (1). Uzyskany nowy podział porównuje się z wcześniejszym podziałem i jeżeli spełnia on warunek kryterium ε, to uzyskany nowy wynik przyjmuje się jako rozwiązanie końcowe. W przeciwnym przypadku powtarza się proces od wyznaczania nowych środków klastrów.

4. Algorytm programu KLASYFIKACJA1

W celu dokonania klasyfikacji zbioru danych zawierającego opis około 500 próbek węgla dokonano modyfikacji algorytmu grupowania wokół c-centroidów. Program ma postać sekwencji podprogramów przedstawionych w postaci bloków. Warto zwrócić uwagę na fakt, że wewnątrz każdego bloku występują skomplikowane operacje na danych w postaci licznych sprzężeń zwrotnych (pętli). Ponadto każdy blok, z wyjątkiem ostatniego, jako wynik generuje dane potrzebne do wykonania operacji w kolejnym bloku. Wobec tego kolejność bloków jest ściśle określona. Oto schemat blokowy zmodyfikowanego algorytmu grupowania:





W poniższych podpunktach omówiono szczegółowo budowę modułów programu.

4.1. Parametry wywołania programu

Program KLASYFIKACJA1 został napisany w środowisku Turbo Pascal. W celu zwiększenia uniwersalności programu do klasyfikowania danych pewne założenia potrzebne do klasyfikacji danych zostały utworzone jako parametry programu. Program wywoływany jest z następującymi parametrami: *plik wejściowy* - zawierający opis danych do klasyfikacji; *plik wyjściowy* - zawierający opis przedstawiające klasyfikację danych; *liczba klas*; *definicja niepodobieństwa danych.*

Algorytmy clusteringu operują na jednej z dwóch struktur danych wejściowych:

a) *p*-wymiarowe obiekty w postaci *p* wartości zmiennych (lub cech atrybutów) ujęte w macierzy *n x p*, gdzie wiersze odpowiadają obiektom, a kolumny zmiennym,

b) zbiór miar sąsiedztwa każdej pary obiektów zbioru danych w postaci macierzy *n x n*, przy czym rozróżnia się dwa typy sąsiedztwa:
   - niepodobieństwo określające, jak różne są dwa obiekty,
   - podobieństwo określające, jak bardzo podobne do siebie są dwa obiekty.

Program KLASYFIAKCJA1 jest wykorzystywany do klasyfikacji próbek węgla. W programie *dane wejściowe* są wartościami zmiennych opisujących próbki, czyli stanowią strukturę typu a). Do analizy wykorzystano zbiór danych zawierający opis próbek węgla

pobieranych na KWK „Knurów". Próbki te opisują wyniki pomiarów co cztery minuty. Do analizy przyjęto pełny zbiór 500 próbek opisanych ośmioma zmiennymi:

x_1 - zawartość (udział wyrażony ułamkiem) frakcji ziaren węgla o gęstości $\rho < 1,3\ g/cm^3$,

x_2 - zawartość frakcji ziaren węgla o gęstości $1,3 < \rho < 1,5\ g/cm^3$,

x_3 - zawartość frakcji ziaren węgla o gęstości $1,5 < \rho < 1,8\ g/cm^3$,

x_4 - zawartość frakcji ziaren węgla o gęstości $\rho > 1,8\ g/cm^3$,

x_5 - zawartość popiołu we frakcji pierwszej o gęstości $\rho < 1,3\ g/cm^3$,

x_6 - zawartość popiołu we frakcji drugiej o gęstości $1,3 < \rho < 1,5\ g/cm^3$,

x_7 - zawartość popiołu we frakcji trzeciej o gęstości $1,5 < \rho < 1,8\ g/cm^3$,

x_8 - zawartość popiołu we frakcji czwartej o gęstości $\rho > 1,8\ g/cm^3$.

Wartości skrajne zostały odrzucane jako obarczone dużymi błędami pomiarowymi (przypisano im wysokie wartości wag). Każdej próbce przypisano czas dokonywanych pomiarów, dzięki czemu można stosować także statyczne metody modelowania oraz obserwować dynamikę zmian właściwości próbek.

W algorytmie klasyfikacji danych o niepodobieństwie próbek, a tym samym o podziale próbek na przedziały decydują tylko wartości dla gęstości $\rho < 1,5\ g/cm^3$.

Istnieją ograniczenia dotyczące *liczby klas*. Liczba ta musi być ze zbioru liczb naturalnych dodatnich i mniejsza niż połowa liczebności próbek węgla znajdujących się w pliku wejściowym. Ograniczenie to wyklucza jako rozwiązanie wymuszenie utworzenia klasy jednoelementowej. Można zastanowić się nad mocniejszym zawężeniem wartości liczby klas podziału próbek, aby zapobiegać tworzeniu się mało liczebnych klas.

O klasyfikacji danych na klasy decyduje stopień niepodobieństwa pomiędzy poszczególnymi próbkami oraz istniejącymi klasami. Dlatego istotna jest przyjęta *definicja niepodobieństwa* będąca miarą w algorytmie grupowania. W programie KLASYFIKACJA1 przyjęto euklidesową miarę niepodobieństwa danych.

4.2. Odczyt danych wejściowych z pliku

Plik danych wejściowych jest plikiem tekstowym. W tym bloku programu następuje konwersja danych z postaci tekstowej do postaci struktury dynamicznej. Zachowana jest liczba zmiennych opisujących każdą próbkę. Odczytywane są kolejne próbki i szeregowane chronologicznie w liście jednokierunkowej zgodnie z kolejnością ich odczytywania z pliku. Wprowadzenie struktur dynamicznych **nie** ogranicza programowo liczby próbek węgla do

sklasyfikowania. Jedyne ograniczenie wynika z wielkości pamięci operacyjnej sprzętu. Dodatkowo próbki są zliczane.

4.3. Normalizacja wartości próbek węgla

Znajdujące się w bazie danych wartości zmiennych opisujących grupowane próbki węgla nie są mierzone w jednolity sposób. W klasyfikacji operującej pierwotnymi wartościami zmiennych zmiana jednostki miary może prowadzić do innego podziału niż podział próbek przed zmiana jednostki. Fleiss i Zubin [5] udowodnili, że brak unormowania wartości zmiennych miewa wpływ na zacieranie się różnic między zmiennymi. Aby wyeliminować zależność wyników podziału od przyjętej jednostki miary, zastosowano normalizację danych do przedziału wartości <0, 1>. W tym celu wykorzystano wartości maksymalne i minimalne zmiennych w zbiorze wszystkich próbek węgla.

Miarę normalizacji określa zależność:

$$\hat{x}_{k,i} = \frac{x_{k,i} - MIN(x_i)}{MAX(x_i) - MIN(x_i)} \tag{2}$$

gdzie:

$x_{i,k}$ - pierwotna wartość *i-tej* zmiennej *k-tej* próbki węgla,

$\hat{x}_{i,k}$ - znormalizowana wartość *i-tej* zmiennej *k-tej* próbki węgla,

$MIN(x_i)$ - wartość minimalna *i-tej* zmiennej w zbiorze wszystkich próbek,

$MAX(x_i)$ - wartość maksymalna *i-tej* zmiennej w zbiorze wszystkich próbek.

Normalizacja jest równoznaczna z przypisaniem *względnych wag* określających stopień, z jakim wartość zmiennej decyduje o niepodobieństwie poszczególnych próbek.

4.4. Wyznaczanie początkowych współrzędnych centroidów

Parametrem algorytmu klasyfikacji danych wokół c-centroidów jest wyznaczony początkowy podział próbek węgla. Zadanie to jest realizowane w dwóch kolejnych blokach programu KLASYFIKACJA1. Podział poprzedza kategoryzacja danych.

Kategoryzacja danych polega na początkowym wyborze zmiennych decydujących o podziale próbek na klasy. Należy ograniczyć liczbę zmiennych, które decydują o klasyfikacji, gdyż nadmiar właściwości może jedynie zaciemnić strukturę klas. Nie istnieją teoretyczne podstawy do określenia optymalnej liczby zmiennych uwzględnianych w analizie. Jest to dobierane eksperymentalnie i zależy od zdefiniowanego celu zadania.

Jedną z metod wyznaczenia podziału początkowego jest losowe wyróżnienie *c* próbek poprzez wskazanie ich jako centroidów początkowych. Metoda ta obarczona jest bardzo dużą przypadkowością. Zmodyfikowany algorytm dąży do uzyskania jak najbardziej równomiernego podziału próbek na klasy z zachowaniem wszystkich warunków klasyfikacji typu *clustering*. Wobec tego stworzono specjalny mechanizm wyboru próbek na centroidy początkowe.

Każdej próbce przypisuje się wyliczoną miarę. Jest to średnia arytmetyczna znormalizowanych wartości zmiennych uwzględnianych przy wyznaczaniu niepodobieństwa między próbkami, czyli x_1, x_2, x_5, x_6.. Następnie zbiór próbek jest sortowany według tej miary. W wyniku sortowania powstaje dynamiczna lista próbek węgla o rosnących wartościach miary. Uzyskaną listę hipotetycznie dzieli się na *c* przedziałów, które są w przybliżeniu równoliczne i znajduje się próbki będące medianami tych przedziałów. Stają się one centroidami początkowymi szukanych klas próbek węgla. Automatycznie są one traktowane jako już sklasyfikowane.

4.5. Wyznaczanie klas węgla poprzez podział próbek na klastry

W tym bloku programu KLASYFIKACJA1 przeprowadzany jest podział zbioru próbek węgla na *c* klastrów. Wyróżnionych wcześniej *c* próbek stanowi podział początkowy. Wartości ich zmiennych pokrywają się z wartościami współrzędnych centroidów. Warto zaznaczyć, że współrzędne centroidu są reprezentacją klasy.

Każda próbka jest porównywana z centroidami kolejnych klas. Stosując wybraną miarę niepodobieństwa oraz na podstawie wartości zmiennych, określa się odległość pomiędzy tą próbką a każdym centroidem. Próbka jest przypisywana do klasy, względem której klasa centroidu jest najbliższa. Po każdym przypisaniu próbki, dla powiększonej klasy modyfikowane są wartości centroidu. Modyfikacja polega na wyznaczeniu nowych średnich wartości współrzędnych centroidu z uwzględnieniem dołączonej próbki.

Na podział próbek wpływ ma także dobór odpowiedniej miary niepodobieństwa próbek. Istnieje wiele sposobów określania miary niepodobieństwa (odległości), które zależą od typu zmiennych.

Niepodobieństwo definiuje się jako nieujemną liczbę *d(i, j)*, przyjmującą tym mniejszą wartość dla obiektów *i-tego* oraz *j-tego*, im bardziej są podobne do siebie te obiekty. Wartość ta rośnie wraz ze zwiększaniem się różnicy między obiektami *i-tym* oraz *j-tym*. Miara ta spełnia założenia:

$$0 \le d(i, j) \le 1 \tag{3.a}$$

$$d(i, i) = 0 \tag{3.b}$$

$$d(i, j) = d(j, i) \tag{3.c}$$

Najczęściej stosowane współczynniki niepodobieństwa to:

$$\sum_{k=1}^{p}(x_{ik} - x_{jk})^2 \qquad \text{odległość euklidesowa} \tag{4.a}$$

$$\sum_{k=1}^{p}\left|x_{ik} - x_{jk}\right| \qquad \text{blok miast zw. Manhattan} \tag{4.b}$$

$$\sum_{k=1}^{p}\frac{\left|x_{ik} - x_{jk}\right|}{\left(x_{ik} + x_{jk}\right)} \qquad \text{miara „Canberra"} \quad \text{dla } x_{ik},\ x_{jk} > 0 \tag{4.c}$$

Uogólnieniem dwóch pierwszych metryk jest *odległość Minkowskiego* zdefiniowana jako:

$$d(i,j) = \left(\left|x_{i1} - x_{j1}\right|^q + \left|x_{i2} - x_{j2}\right|^q + \ldots + \left|x_{ip} - x_{jp}\right|^q\right)^{1/q/q} \tag{4.d}$$

gdzie q - liczba rzeczywista większa lub równa 1.

W programie KLASYFIKACJA1 wykorzystuje się odległość euklidesową. Aby jej wartości nie zależały od skal przedziałów zmiennych, dokonuje się normalizacji zmiennych prowadzącej do wyznaczenia wspólnego przedziału wartości zmiennych. Operacja ta była omawiana we wcześniejszej części tego referatu.

4.6. Autokorekta podziału próbek

Celem klasyfikacji jest stworzenie modelu statystycznego jak najlepiej odzwierciedlającego dane rzeczywiste i nie narzucającego sztucznych struktur odbiegających od fizycznej interpretacji. Dlatego wyznaczony podział poddawany jest testowaniu, podczas którego jest on ulepszany poprzez mechanizm samouczenia się.

Na tym etapie wykonywania programu próbki węgla zostały podzielone na klasy. Każda klasa określona jest przez centroid, którego wartości współrzędnych są średnimi arytmetycznymi wartości poszczególnych zmiennych próbek należących do wskazywanego klastra. Jest to pierwszy proponowany podział jako rozwiązanie i należy poddać go korekcie. Jest to przykład zastosowania mechanizmu autokorekty, czyli mechanizmu samouczącego się.

Korekta jest potrzebna, gdyż z dużym prawdopodobieństwem istnieją próbki, które zostały błędnie przypisane do klas. Wynika to z faktu, że po przyporządkowaniu próbki do klastra wyznaczane są nowe, aktualne, współrzędne centroidu. Na etapie pierwszej

klasyfikacji celowo zaniechano sprawdzenia zgodności przyporządkowania pozostałych próbek już istniejących w klasie względem centroidu po aktualizacji jego wartości. Wyodrębnienie czynności autokorekty w postaci odrębnego bloku wynika z optymalizacji czasowej programu. Każdorazowe korygowanie po sklasyfikowaniu pojedynczej próbki podniosłoby złożoność obliczeniową wcześniejszego bloku. Wynika to z wystąpienia wielu operacji pustych, tzn. stwierdzania poprawnego przyporządkowania próbek do klas.

W programie KLASYFIKACJA1 procedura autokorekty polega na wykrywaniu przypadków przypisania próbki węgla do klasy, względem której miara niepodobieństwa nie jest minimalna. W ramach autokorekty próbka taka jest przemieszczana do klasy, względem której ma ona miarę niepodobieństwa minimalną. Następnie aktualizuje się współrzędne centroidów klas, które zostały zmienione, czyli: klasy, z której pobrano próbkę oraz klasy, do której dołączono próbkę. Procedura autokorekty jest wywoływana w sposób rekurencyjny ponownie dla całego zbioru próbek. Wykonywanie procedury się kończy w momencie stwierdzenia poprawnego przypisania każdej próbki węgla.

4.7. Wyznaczenie rozmytych współczynników przynależności próbek

Metody grupowania rozmytego wykorzystują rozmytość wprowadzając stopnie prawdziwości stwierdzenia, aby uniknąć podejmowania „ostrych" decyzji. Celem analizy grupowania rozmytego (*ang.: fuzzy clustering*) jest tworzeniem podziałów, w których każdy obiekt przynależy w różnym stopniu do wszystkich klastrów.

Podczas grupowania rozmytego (*ang.: fuzzy clustering*) tworzone są podziały P_f, w których każdy obiekt przynależy w różnym stopniu do wszystkich klastrów. Stopień ten określają współczynniki przynależności (*ang.: membership coefficients*), które spełniają następujące warunki:

$$1.\ \mu_{ik} \in [0, 1] \qquad 1 \le i \le c, \qquad 1 \le k \le n \tag{5.a}$$

$$2.\ \sum_{i=1}^{c} \mu_{ik} = 1 \qquad 1 \le k \le n \tag{5.b}$$

$$3.\ 0 < \sum_{k=1}^{n} \mu_{ik} < n \quad 1 \le i \le c \tag{5.c}$$

W programie KLASYFIKACJA1 procedura wyznaczania współczynników rozmycia dla każdej próbki węgla względem każdej klasy połączona jest z rekonfiguracją bazy danych. Tworzona jest nowa struktura dynamiczna, w której rekordy zawierają numer próbki oraz odpowiadające tej próbce współczynniki rozmycia.

Rozmyty współczynnik przynależności jest funkcją odległości między badaną próbką a kolejnymi centroidami. Współczynnik przynależności *i-tej* próbki węgla do *k-tego klastra* wyliczany jest według zależności:

$$\mu_{ik} = \frac{1 - d(x_i - v_k)}{LKlas - \sum_{i=1..n, j=1..c} d(x_i - v_j)} \tag{6}$$

gdzie:

$d(x_i, v_k)$ - odległość pomiędzy *i-tą* próbką węgla a *k-tym* klastrem,

v_k - centroid *k-tego* klastra.

4.8. Wyznaczenie miar oceny rozmytości podziału

W celu prezentacji graficznej uzyskanych podziałów danych dokonuje się wyostrzenia podziału (defuzyfikację).Wyostrzenie podziału to zamiana klasyfikacji rozmytej na klasyfikację typu crisp. Proces wyostrzania opiera się na macierzy przynależności podziału. Polega on na zastąpieniu wartości funkcji przynależności obiektów u_{iv} na wartości w_{iv} w następujący sposób:

$w_{iq} = 1$ dla *v-tego* klastra, względem którego obiekt ma największą wartość współczynnika przynależności,

$w_{iq} = 0$ dla pozostałych klastrów.

Do miar oceny rozmytości grupowania, czyli stopnia różnicy między otrzymanym rozwiązaniem rozmytym a podziałem typu crisp należą między innymi:

1. **Współczynniki podziału** (ang.: partition coefficient) $F(\tilde{U}; c)$
2. **Entropia podziału** (ang.: partition entropy) $H(\tilde{U}; c)$

Definicja. Niech $\tilde{U} \in M_{fc}$ jest rozmytym podziałem n obiektów w *c* klastrach. ***Współczynnik podziału***, zwany współczynnikiem podziału Dunna $\tilde{U}$ [3], jest skalarem obliczanym w następujący sposób:

$$F(\tilde{U}; c) = \sum_{k=1}^{n} \sum_{i=1}^{c} \frac{(\mu_{ik})^2}{n} \tag{7}$$

gdzie U - macierz wszystkich przynależności.

Dla całkowitego rozmycia, czyli dla $\mu_{iv} = 1/k$, współczynnik Dunna przyjmuje wartość minimalną $F_k = 1/k$, a dla podziału ostrego - wartość maksymalną $F_k = 1$.

Definicja. ***Entropia podziału*** rozmytego $\tilde{U} \in M_{fc}$ zbioru X w c klastrach, gdzie $|X| = n$ oraz $1 \le c \le n$, jest obliczana według zależności:

$$H(\tilde{U}\,;\,c) = \frac{1}{n}\sum_{k=1}^{n}\sum_{i=1}^{c}\mu_{ik}\ln(\mu_{ik}) \qquad (8)$$

Dla całkowitego rozmycia, czyli dla $\mu_{ik} = 1/k$, entropia przyjmuje wartość minimalną $H(\tilde{U}\,;\,c) = 0$, a dla podziału ostrego - wartość maksymalną $H(\tilde{U}\,;\,c) = \ln(c)$.

W programie KLASYFIKACJA1 jako miary rozmycia uzyskanego podziału końcowego obliczane są: współczynnik podziału oraz entropia podziału.

4.9. Zapis danych wyjściowych

Program KLASYFIKACJA1 jako rozwiązanie generuje plik wyjściowy zawierający opis uzyskanych podziałów.

Plik wyjściowy tworzony przez program zawiera opis uzyskanego podziału końcowego. Znajdują się w nim:

- współczynniki określające stopień rozmycia podziału końcowego: współczynnik podziału Dunna oraz entropia podziału;
- opis kolejnych klas; opis każdej klasy obejmuje wartości zmiennych centroidu tej klasy oraz wyszczególnienie próbek węgla należących do tej klasy przy podziale ostrym; przy próbkach węgla podane są wartości zmiennych x_1, x_2;
- opis rozmytego podziału wszystkich próbek węgla wraz z numerem próbki podane są wartości współczynników przynależności tej próbki do każdej z klas.

Wraz z zapisem danych do pliku zwalniana jest pamięć operacyjna poprzez usuwanie struktur dynamicznych.

5. Wnioski

Metody analizy typu clustering są cennym narzędziem w badaniu danych wielu zmiennych. Organizując dane w klastrach można odkryć zależność występującą między zmiennymi obiektów lub wyznaczyć wzorce danych.

Przedstawiony w referacie zmodyfikowany algorytm grupowania danych wokół c-centroidów dzieli zbiór próbek węgla na podaną parametrycznie liczbę klas. Kryteriami rozmycia podziału są obliczone współczynnik podziału Dunna oraz entropia podziału. Jako rozwiązanie program wskazuje optymalny podział próbek węgla pod względem podobieństwa próbek wewnątrz klas. Podczas analizy dzielącej zbiór zawierający opis 500 próbek węgla na żądaną liczbę klas wraz ze wzrostem liczby klas poprawia się kryterium podziału. Ze względu

na ograniczoną wielkość referatu analiza uzyskanych wyników klasyfikacji zostanie przedstawionej w kolejnej publikacji.

LITERATURA

1. Bezdek J. C.: Pattern recognition with fuzzy objective function algorithms. Morton Nadler. 1981.
2. Bezdek J.C.: Fuzzy models and digital signal processing (for pattern recognition): is this a good marriage? - Digital Signal Processing, 3. 1993, str. 253 - 270.
3. Dunn J.C.: A graph-theoretic analysis of pattern classification via Tamura's fuzzy relation, IEEE Trans. on Systems, Man and Cybernetics SMC-4, 1974, str. 310-313.
4. Everitt Brian S.: Cluster Analysis. Americas, Halsted Press, New York, 1993.
5. Fleiss J. L., Zubin J.: On the methods and theory of clustering. Multivariate Behavioral Research, 4. 169, str. 235-250.
6. Gower J.C.: Maximal predictive classification. Biometrics, 30. 1974, str. 643-654.
7. Kowal A.: Zastosowanie logiki rozmytej do klasyfikacji danych. VI Konferencja Automatyzacji Procesów Przeróbki Kopalin. Szczyrk. V 2000 r., str. 85-106.
8. Kowal A.: Optymalizacja klasyfikacji próbek węgla z wykorzystaniem rozmytej metody grupowania (c-centroidów). XIV Krajowa Konferencja Automatyki. Zielona Góra. VI 2002 r.
9. Walaszek-Babiszewska A., Kowal A.: Wykorzystanie grupowania rozmytego do klasyfikacji próbek węgla. Międzynarodowa konferencja „Górnictwo 2000", Szczyrk 1999, str. 169 – 179.

Recenzent: Dr hab. inż. Anna Walaszek-Babiszewska, prof. Uniw. Zielon.

Abstract

In the article modified c-means clustering algorithm of coal samples is presented. In the second part basic features of crisp clustering methods are mentioned. In the third part of the article general distribution of the fuzzy pattern recognition methods is discussed. Also classical c-means clustering algorithm is shown.

In the next parts, a scheme of modified clustering algorithm is introduced and the following procedures are presented: entering of input parameters of the program; reading input data from the file; samples' values normalisation; samples classification accord to average measure; calculation of primary centroid's coordinates; samples' grouping – clustering; auto-correction of samples' partitions; calculation of the fuzzy membership coefficients; calculation of the measures of partition's fuzziness; recording output data.