**Temat 1 i Temat 2**

**Temat 1 Grupowanie wartości atrybutów obiektów treningowych w celu przygotowania danych do tworzenia pokryć, które będą wykorzystane w modelu klasyfikacji**

**Temat 2 Tworzenie pokryć zbioru danych treningowych oraz wyznaczanie otoczeń obiektów treningowych w celu wykorzystania w modelu klasyfikacji**

**Temat wspólny**

**Grupowanie wartości atrybutów obiektów treningowych i tworzenie pokryć tego zbioru wraz z funkcją oceny oraz wyznaczanie otoczeń obiektów treningowych w celu wykorzystania w modelu klasyfikacji**

**Oznaczenia:**

**U-zbiór obiektów treningowych (u1, u2 … un )**

**T - zbiór obiektów testowych (t1, t2 … tm )**

**Atr – zbiór atrybutów (a1,a2… )**

**D – atrybut decyzyjny**

**Val(a) – zbiór wartości atrybutu a**

**n(a) = moc zbioru Val(a)**

**Pos(a)- zbiór obiektów dla których wartość atrybutu decyzyjnego jest równa 1 oraz wartość atrybutu a jest określona**

**P(a) =Pos(a)/n(a) – frakcja pozytywnych decyzji względem atrybutu a**

1. Obliczenia wykonujemy dla każdego atrybutu oddzielnie
   1. Grupujemy dane w klasy kl1, kl2, … Niech k oznacza liczbę klas.
   2. Grupowanie
      1. Dane dyskretne -kazda klasa wyznaczona jest jako wartość atrybutu (w przypadku, gdy n(a) >20 możliwe jest jakieś pogrupowanie)
      2. Dane ciągłe (Real) – grupujemy w szereg rozdzielczy
         1. Różny podział na klasy jak na zajęciach MMS
         2. Podział na klasy przy użyciu kwantyli
      3. Inne metody grupowania
   3. Obliczenia
      1. Dla każdej klasy ki obliczamy
         1. Liczebność klasy ni
         2. Liczba pozytywnych decyzji w klasie nPosi
         3. Warunkowa frakcja jedynek pPosi=nPosi/ni
         4. factori=pPosi/P(a)
2. Wyznaczenie pokryć (**Uwaga: to będzie część wspólna z tematem2).**  Pokrycia tworzymy otrzymując dla każdego atrybutu dwa zbiory pokrywające LOWa oraz HIGHa otrzymane na podstawie dwóch parametrów 0<Low<1, 1<High<2 (dla różnych wartości parametrów będą różne pokrycia, to będzie podlegało ocenie zob. Ocena)
   1. Dla ustalonego atrybutu a i dla każdej klasy kli wyznaczamy
      1. LOW(kli)=1 jeżeli factori<Low , a w p.p. jest równy 0
      2. HIGH(kli)=1 jeżeli factori>High, a w p.p. jest równy 0
   2. Dla każdego obiektu u wyznaczamy LOWa(u)=LOW(kli), gdzie kli jest klasa do której należy u. Jeżeli u ma nieokreśloną wartość atrybutu a, to LOWa(u)=0.
   3. Analogicznie, dla każdego obiektu u wyznaczamy HIGHa(u)=HIGH(kli), gdzie kli jest klasa do której należy u. Jeżeli u ma nieokreśloną wartość atrybutu a, to HIGHa(u)=0.
3. Redukcja tablicy pokryć Cov
   1. Usuwanie kolumn zerowych
   2. Usuwanie kolumn jedynkowych
4. Ocena – dla każdego atrybutu i każdego obiektu testowego t oceniamy, czy t jest poprawnie klasyfikowane tzn.:
   1. Jeżeli D(t)=1 i LOWa(t)=1, to liczymy jako błąd
   2. Podobnie, jeżeli D(t)=0 i HIGHa(t)=1, to liczymy jako błąd
   3. Zliczamy błędy dla wszystkich atrybutów.
5. Wykorzystanie powyższej Oceny :
   1. Do oceny sposobu grupowania przy ustalonych Low i High (temat1)
   2. Do wyznaczenia najlepszych Low i High przy ustalonym sposobie grupowania (temat 2)
6. Wyznaczanie otoczeń obiektów testowy polega na obliczeniu dla każdego obiektu testowego t (dla każdej niezredukowanej kolumny) LOWa(t) i HIGHa(t)