Algorytm rangowania danych

Spis treści

1. RANKING KONDYCJI ZAWODÓW

System Blender Danych daje możliwość wygenerowania rankingu kondycji zawodów, który:

- pozwala spojrzeć na zawody z perspektywy kilku czynników łącznie – uszeregować je w oparciu o zestaw wskaźników pozwalających ocenić je w kontekście sytuacji na rynku wakatów, napływu zatrudnionych, bezrobocia rejestrowanego oraz wypłacanych wynagrodzeń;
- przygotowywany jest na poziomie województw i kraju;
- jest tworzony przy użyciu algorytmu rangowania danych opartym na metodzie TOPSIS;
- zawody są rangowane w wyniku ich porównania w oparciu o wybrane wskaźniki – opisane w pliku „5_RANKING KONDYCJI ZAWODÓW – Wskaźniki" (w zakładce Metodyka Blendera Danych)

2. OPIS METODYCZNY ALGORYTMU RANGOWANIA DANYCH

a) Założenia algorytmu

Model algorytmu przetwarzania i rangowania danych w systemie Blender Danych zaprojektowano przyjmując następujące założenia:

- produktem algorytmu ma być ranking zawodów przygotowany na podstawie wielu zmiennych. Powinien odzwierciedlać zarówno wymiar podażowy, jak i popytowy pozycji danego zawodu oraz ocenę wynikającą z tego porównania. Powinna być dostępna możliwość prezentacji zarówno cząstkowych zestawień (poziom województwa), jak i raportów zagregowanych (poziom kilku połączonych województw, do skali kraju);
- algorytm powinien pozwolić na zbudowanie rankingu niezależnie od typu danych wejściowych – musi zatem zawierać procedury normalizacyjne;
- algorytm powinien umożliwiać nadawanie wag poszczególnym zmiennym wchodzącym w skład rankingu, z uwzględnieniem metod normalizacji danych adekwatnych dla różnych procedur;
- algorytm powinien umożliwiać wykorzystanie różnych procedur metodycznych wyliczania rankingu, zapewnić możliwość podejścia multimetodycznego, a więc bazującego na różnych procedurach rangowania.

Algorytm powinien być możliwie zautomatyzowany. Oznacza to zaszycie procedur analitycznych algorytmu w oprogramowaniu. Poszczególne procedury powinny być dostępne

przynajmniej z poziomu operatora systemu Blendera Danych w postaci prostego pola wyboru. Wybór danej procedury powinien prowadzić do wygenerowania całej analizy na podstawie dostarczonej bazy danych wejściowych.
Algorytm powinien umożliwiać analizę wpływu czynników społeczno-gospodarczych na sytuację danego zawodu. Czynniki te mogą mieć różnorodny charakter (odnosić się do m.in. stanu bezrobocia, stanu zatrudnienia). Ocena zawodów pod kątem związku z oddziaływaniem tych czynników daje potencjalną możliwość dodatkowej oceny, przez co może stanowić dodatkową, istotną informację zarządczą.

b) Operacje wejściowe

Wejściowym zbiorem danych jest macierz (tablica) danych charakteryzujących zawody

$$X = [x_{ij}]$$

gdzie:
- *i – numer wiersza (przypadku) indeks numeru zawodu (i=1,…,m);*
- *j – indeks (numer) zmiennej diagnostycznej charakteryzującej zawody (j=1,…,n)*

Każda ze zmiennych w bazie (tablicy) X powinna mieć możliwość przypisania jej charakteru (flagi):
- **stymulanta** (im wyższe wartości tej zmiennej dla danego przypadku, tym dany przypadek jest lepszy w badaniu – rankingu – ze względu tylko na tę zmienną)
- **destymulanta** (im niższe wartości tej zmiennej dla danego przypadku, tym dany przypadek jest lepszy w badaniu – rankingu – ze względu na tę zmienną)
- **nominanta** (im bliższe wartości zmiennej pewnej wartości nominalnej punktowej c_{0j} lub z pewnego przedziału $[c_{1j}, c_{2j}]$, tym dany przypadek jest lepszy w badaniu – rankingu ze względu na tę zmienną)

Dla zmiennych o charakterze nominant z bazy powinna być możliwość ustalenia i przypisania wartości nominalnych w postaci zbioru zamkniętego.

Wybór zmiennych do rankingu z bazy (tablicy) X: (wybierając zmienne program powinien informować o charakterze zmiennej i podawać ewentualne wartości nominalne)
- pod względem popytowym (pracodawcy, zakłady pracy, firmy),
- pod względem podażowym (np. absolwenci uczelni, szkół średnich),
- wszystkie dostępne zmienne z bazy.

Ustalanie wag (ważności) dla zmiennych – powinna być dostępna opcja:
- nadania takich samych wag dla wszystkich zmiennych: $\omega_j = \frac{1}{n}$,

- w wariancie gdy wagi są takie same – można je pominąć,
- ewentualnie przypisania odpowiednich wag dla każdej indywidualnej wybranej zmiennej diagnostycznej

$$0 < \omega_j < 1; \sum_{j=1}^{n} \omega_i = 1 \tag{1}$$

3. REALIZACJA PROCEDURY RANKINGU METODĄ TOPSIS

3.1. Ustalenie charakteru zmiennych

W metodzie TOPSIS zakłada się, że wszystkie wybrane zmienne diagnostyczne **powinny mieć charakter stymulant lub destymulant**. Zatem te zmienne, które są nominantami należy zamienić na odpowiadające im zmienne, które mają charakter stymulant. Do zamiany można zastosować jedno z następujących przekształceń:

- formułę ilorazową (tylko dla nominant mierzonych na skali ilorazowej o wartościach będących wskaźnikami ilorazowymi o wartościach >=0) przekształca na stymulanty o wartościach z przedziału [0,1]

$$x_{ij} = \frac{\min_i\{c_{oj}; x_{ij}^N\}}{\max_i\{c_{oj}; x_{ij}^N\}} \tag{2}$$

i=1,…,m; j=1,…,n; zaś c_{0j} - są nominalnymi najlepszymi wartościami zmiennych będących nominantami x_{ij}^N.

- formułę różnicową (stymulanta uzyskana tym wzorem jest mierzona na skali przedziałowej)

$$x_{ij} = -\left|x_{ij}^N - c_{0j}\right| \tag{3}$$

i=1,…,m; j=1,…,n; zaś c_{0j} – są nominalnymi najlepszymi wartościami zmiennych będących nominantami x_{ij}^N.

- można zastosować też formułę:

$$x_{ij} = \begin{cases} \frac{-1}{x_{ij}^N - c_{oj} - 1} & gdy \quad x_{ij} < c_{0j} \\ 1 & gdy \quad x_{ij} = c_{0j} \\ \frac{1}{x_{ij}^N - c_{oj} + 1} & gdy \quad x_{ij} > c_{0j} \end{cases} \tag{4}$$

i=1,…,m; j=1…,n

3.2. Unormowanie (normalizacja) wybranych zmiennych diagnostycznych

Do wyznaczenia wartości unormowanych na skali przedziałowej zastosowano metodę standaryzacji.

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j} \tag{5}$$

gdzie:

i=1,...,m; j=1...,n

$$\bar{x}_j = \frac{\sum_{i=1}^{m} x_{ij}}{m} \tag{6}$$

jest średnią wartością j-tej zmiennej diagnostycznej x_j; m – liczba przypadków (wartości zmiennej); n - liczba wybranych zmiennych do analizy

$$S_j = \sqrt{\frac{\sum_{i=1}^{m} (x_{ij} - \bar{x}_j)^2}{m}} \tag{7}$$

jest odchyleniem standardowym j-tej zmiennej diagnostycznej x_j; m – liczba przypadków (wartości zmiennej)

Każdy obiekt badany (zawód) charakteryzowany wybranymi zmiennymi diagnostycznymi będzie interpretowany jako punkt z przestrzeni wielowymiarowej R^n. Zachodzi zatem konieczność wyboru na potrzeby rankingu odpowiedniej metryki odległości, za pomocą której będą porównywane obiekty.

3.3. Wybór odległości pomiędzy obiektami w przestrzeni wielowymiarowej potrzebnej do wyznaczenia rankingu.

W procedurze obliczeniowej zastosowano miarę klasycznej odległości Euklidesowej:

- **odległość euklidesowa (wagi jednakowe - brak wag)** pomiędzy k-tym oraz p-tym obiektem punktem w przestrzeni wielowymiarowej.

$$d_{k,p} = \sqrt{\left(\frac{1}{n}\right)^2 \sum_{j=1}^{n} (z_{kj} - z_{pj})^2} \quad \text{lub} \quad d_{k,p} = \sqrt{\sum_{j=1}^{n} (z_{kj} - z_{pj})^2} \tag{8}$$

- **odległość euklidesowa (wagi zmienne)** pomiędzy k-tym oraz p-tym obiektem punktem w przestrzeni wielowymiarowej.

$$d_{k,p} = \sqrt{\sum_{j=1}^{n} \omega_j^2 \cdot (z_{kj} - z_{pj})^2} \tag{9}$$

gdzie: $0 < \omega_j < 1, \sum_{j=1}^{n} \omega_j = 1$ – *ustalony system wag dla j-tej zmiennej*

3.4. Budowa rankingu

Na początku wyznacza się abstrakcyjny obiekt najlepszy (idealny) o najlepszych wartościach zmiennych diagnostycznych (maksymalnych dla stymulant i minimalnych dla destymulant).

$$z_0 = [z_{01} \quad z_{02} \quad \dots \quad z_{0n}] \tag{10}$$

$$z_{0j} = \begin{cases} \max_i z_{ij} & gdy \quad Z_j - stymulanta \\ \min_i z_{ij} & gdy \quad Z_j - destymulanta \end{cases}$$

Kolejno wyznacza się drugi obiekt antyidealny w przestrzeni wielowymiarowej (antywzorzec), który jest obiektem o najgorszych wartościach zmiennych diagnostycznych (minimalnych dla stymulant oraz maksymalnych dla destymulant).

$$z_{-0} = [z_{01} \quad z_{02} \quad \dots \quad z_{0n}] \tag{11}$$

$$z_{-0j} = \begin{cases} \min_i z_{ij} & gdy \quad Z_j - stymulanta \\ \max_i z_{ij} & gdy \quad Z_j - destymulanta \end{cases}$$

Następnie bada się podobieństwo obiektów (badanych zawodów) do abstrakcyjnego obiektu najlepszego wyznaczając odległości obiektów od wzorca $d_{i,0}$ – jako miarę należy zastosować przyjętą wcześniej miarę odległości Euklidesowej.

Bada się także podobieństwo analizowanych obiektów do abstrakcyjnego obiektu najgorszego, wyznaczając odległości obiektów od antywzorca $d_{i,-0}$ – jako miarę też należy zastosować przyjętą wcześniej miarę odległości Euklidesowej.

Kolejno wyznacza się wartości dla zmiennej agregatowej (syntetycznej) ze wzoru:

$$R_i(TOPSIS) = \frac{d_{i,-0}}{d_{i,-0} + d_{i,0}} \tag{12}$$

Wartości tej zmiennej syntetycznej należą do przedziału [0,1]. Im wyższe jej wartości dla danego obiektu, tym wyższe jego miejsce w rankingu.

3.5. Wyznaczenie ostatecznych rankingów (posortowanie obiektów nierosnąco) i przypisanie rankingów porządkowych.