Uniwersytet Śląski
Wydział Informatyki i Nauki o Materialach
Kierunek Informatyka
Studia niestacjonarne II stopnia

Praca dyplomowa magisterska

Inteligencja stadna w eksploracji danych

Promotor:
dr hab. Urszula Boryczka

Autor:
Patrycja Tkocz

Sosnowiec 2016

Spis treści

Wstęp

wstęp

1. Eksploracja danych

Eksploracja danych inaczej data mining to jeden z procesów odkrywania wiedzy, wykorzystujący szybkość istniejących maszyn do poszukiwania pewnych zależności, wspólnych cech. Dzięki eksploracji jesteśmy wstanie przeszukać duże zbiory danych uzyskując pewne ukryte w danych cechy, zależności. Jest to proces analityczny, zazwyczaj powiązany z zachowaniami rynkowymi, gospodarczymi. Uzyskane wzorce stosowane są do nowych podzbiorów. Wykorzystywana jest w bazach danych, jak i do rozwiązywania problemów np. poszukiwanie odpowiedzi jaka będzie jutro pogoda w oparciu o pewne dane. Zgłębianie danych pozwala nam na analizę danych w celu ich lepszego zrozumienia.

Metody eksploracji danych można sklasyfikować na wiele sposobów. Zazwyczaj dzielimy je ze względu na cel eksploracji, typy eksplorowanych danych lub typ wzorców odkrywanych w procesie eksploracji. Najpopularniejsza jest klasyfikacja ze względu na cel eksploracji. Tak więc wyróżniamy:

- klasyfikacja
- wyszukiwanie asocjacji
- analiza sekwencji i przebiegów czasowych
- odkrywanie charakterystyk
- eksploracja WWW
- wykrywanie zmian i odchyleń
- grupowanie

Klasyfikacja to metoda analizy danych, której celem jest określenie sposobu przynależności obiektu do pewnych kategorii w zależności od wartości atrybutów, liczebności. Mówiąc inaczej klasyfikacja odpowiada za przypisanie obiektu do jednej z predefiniowanych klas w oparciu o zbiór atrybutów opisujących dany obiekt. Klasyfikacja znalazła szerokie zastosowanie w bankowości - procedura przyznawania kredytu w zależności od charakterystyki konsumenta, medycynie - kwalifikacja pacjentów do zabiegu w zależności od stanu zdrowia i wyników badań pacjenta. Najpopularniejszą metodą w klasyfikacji jest odkrywanie modeli nazywanych klasyfikatorami. Dzięki wykrytym modelom/funkcją możliwe jest wyszukanie nowych obiektów o nieznanej kwalifikacji. Przykładami klasyfikacji mogą być drzewa decyzyjne, modele Bayes'a, sieci neuronowe, k-najbliższych sąsiadów, algorytmy genetyczne. Klasyfikacja może być dokonywana na obiektach z ciągłymi danymi lub też danymi kategorycznymi, sekwencjami danych, danych tekstowych, utworach muzycznych, strukturach grafowych.

Wyszukiwanie asocjacji to jedna z popularnych metod przetwarzania danych. Celem jest znalezienie interesujących asocjacji miedzy danymi w dużych zbiorach, czyli zależności lub korelacji między danymi. W wyniku okrywania asocjacji otrzymujemy zbiór asocjacji nazywany zbiorem reguł asocjacyjnych. Odkrywanie asocjacji znalazło wiele popularnych zastosowań. Przykładem może być stosowanie asocjacji do analizy koszyka zakupów, czyli znalezienia naturalnych wzorców zachowań konsumenta, organizacja akcji promocyjnych. Wykorzystywana jest również w analizie dokumentów, analizie sekwencji DNA, sekwencjach białkowych. Metoda ta jest szeroko wykorzystywana w innych metodach eksploracji danych - klasyfikacja, predykcja, grupowanie.

Analiza sekwencji i przebiegów czasowych obejmuje metody analizy sekwencji danych kategorycznych i przebiegów czasowych. Celem metody analizy sekwencji jest znalezienie podsekwencji czyli wzorców sekwencji, klasyfikacja i grupowanie sekwencji. Natomiast metody analizy przebiegów czasowych dążą do znalezienia trendów, podobieństw, anomalii, a także cykli w przebiegach czasowych.

Odkrywanie charakterystyk to metoda, której celem jest znalezienie charakterystyk (zwięzłych opisów) danego zbioru danych. Odnajdywanie charakterystyk może odbywać się na dwa sposoby. Jednym z sposobów jest odkrywanie charakterystyki zbioru danych, którego celem jest podsumowanie danych należących do podanego zbioru. Poddawany analizie zbiór zazwyczaj jest pobierany z bazy danych

lub hurtowni danych poprzez zapytanie. Po uzyskaniu zbioru, poddajemy go dalszej analizie w celu wyszukania charakterystyki tego zbioru. Drugim ze sposobów jest analiza dyskryminacyjna zbioru danych, która polega na porównaniu podstawowych cech zbioru z cechami zbioru porównawczego. W tym sposobie dane dwóch zbiorów również pochodzą z zapytania do bazy bądź hurtowni danych. Po ich otrzymaniu następuje ich porównanie i analiza. Wyniki odkrywania charakterystyk w obu metodach przedstawiane są w postaci wykresów graficznych, reguł charakterystycznych dla sposobu 1 lub reguł dyskryminacyjnych dla sposobu 2.

Eksploracja WWW to w ostatnim czasie jedna z najprężniej rozwijających się metod eksploracji danych, co wynika z rozwoju sieci Web. Metoda ta zajmuje się odkrywaniem nieznanej dotąd wiedzy czyli reguł, wzorców i zależności ukrytych w zawartości sieci Web i sposobie korzystania z niej. Mówiąc prościej obejmuje analizę korzystania z WWW w celu znalezienia typowych wzorców zachowań użytkowników w sieci. Sieć Web jest pewnego rodzaju baza danych ale dane nie są przechowywane w sposób strukturalny, a także cechują się dużą złożonością. Przechowywane są w w logach serwerów WWW, mając duże rozmiary i dynamiczny przyrost. Na potrzeby eksploracji sieci WWW stworzono wiele nowych metod eksploracji danych. Zazwyczaj wyróżnia się 3 podstawowe grupy tj. eksploracja zawartości sieci Web, eksploracja połączeń sieci Web, eksploracja korzystania z sieci Web. Eksploracja WWW znalazła zastosowanie przy wspomaganiu działania wyszukiwarek sieciowych, grupowaniu i klasyfikacji stron WWW, handlu elektronicznym, reklamach internetowych, optymalizacji działania systemów baz danych.

Wykrywanie zmian i odchyleń obejmuje metody analizy danych zmiennych w czasie, których celem jest znalezienie różnic pomiędzy aktualnymi a oczekującymi wartościami danych. Uściślając metody te odpowiadają za wyszukanie anomalnych zachowań klientów np. w ubezpieczalni, bankomacie.

Grupowanie to również jedna z popularnych metod eksploracji danych. Polega na grupowaniu obiektów o podobnych cechach w klasy, które nazywamy klastrami lub skupieniami.Istniej wiele technik grupowania, a najpopularniejsze z nich to grupowanie hierarchiczne i oparte na podziele. Grupowanie znalazło szerokie zastosowanie w bankowości, medycynie, przetwarzaniu tekstów. Więcej informacji znajduje się w kolejnym podrozdziale.

Podsumowując eksploracja danych to proces automatycznego odkrywania wzor-

ców, reguł, zależności, podobieństw i trendów w repozytoriach danych. Podstawowym jej celem jest analiza danych i procesów w celu lepszego zrozumienia pewnej ilości danych. Jest to bardzo rozległa dziedzina z którą każdy człowiek ma do czynienia, niekoniecznie będąc tego świadomym.

1.1. Grupowanie danych

Grupowanie jest jednym z możliwych podrozdziałów eksploracji danych, jest to proces wyszukiwania wzorców, znalezienia podobieństwa pomiędzy obiektami, a także wskazania wspólnych cech. Ogólnie pojęcie grupowanie danych data clustering to zagadnienie polegające na podziale wejściowego zbioru na mniejsze grupy, gdzie osobniki najbardziej do siebie podobne powinny znajdować się w tej samej grupie, natomiast osobniki różniące się w innych grupach. Ten typ eksploracji stosujemy w celu prawidłowego rozwiązania problemu gdzie mamy dany zbiór obiektów i problemem jest znalezienie naturalnego pogrupowania obiektów w klasy nazywane klastrami lub skupieniami, przy założeniu, że klasy różnią się od siebie znacząco, a osobniki w danym klastrze są do siebie zbliżone lub posiadają te same cechy. Cel grupaowanie możemy zobrazowac na Rys.1

Analizy skupień możemy dokonać na podstawie różnych technik [1]. Wyróżniamy:

- grupowanie oparte na podziele
- grupowanie hierarchiczne
- grupowanie oparte na gęstościach

Grupowanie oparte na podziele - inaczej nazywane grupowaniem optymalizacyjno - iteracyjnym to metoda pozwalająca podzielić zbiór danych na k skupień, gdzie k jest parametrem wywołania takiego grupowania. Wykorzystują one kryterium minimalizacji funkcji to znaczy, że o przynależności do klastra k_i decyduje najmniejsza wartość funkcji.

Precyzując jeśli mamy dany zbiór danych Z który chcemy pogrupować na k rozłącznych zbiorów tj. $C = \{C_1, ..., C_k\}$, gdzie każdy elementy $p \in Z$ przynależy tylko i wyłącznie do jednego skupienia C. Każdy klaster posiada "środek ciężkości"m_k, dzięki któremu jesteśmy wstanie określić przynależność elementu p do któregoś ze

klastrów C. Aby tego dokonać obliczana jest odległość danych od "środka"klastra. W tego rodzaju grupowaniu mamy do czynienia z kilkoma funkcjami odległości:

- odległość euklidesowa

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}, \tag{1}$$

- odległość miejska

$$d(x,y) = \sqrt{\sum_{i=1}^{n}|x_i - y_i|}, \tag{2}$$

- ważona odległość euklidesowa

$$d(x,y) = \sqrt{\sum_{i=1}^{n}w_i(x_i - y_i)^2}, \tag{3}$$

- ważona odległość miejska

$$d(x,y) = \sqrt{\sum_{i=1}^{n}w_i|x_i - y_i|}, \tag{4}$$

O jakości grupowania najczęściej decyduje funkcja

$$E(C) = \sum_{i=k}^{K}\sum_{p\in C_k} d^2(p, m_k) \tag{5}$$

zwana sumą błędów kwadratowych, gdzie d oznacza jedną z funkcji odległości.

Przykładem takiej analizy skupień są algorytm K-średnich oraz K-mediana.

Grupowanie hierarchiczne to grupowanie oparte o tworzenie "drzew". Liście to obiekty, a węzły to klastry. W zależności od kierunku hierarchii rozróżniane są dwie metody:

- aglomeracyjną - polega na łączeniu punktów w grupy;
  Schemat działania takiej metody jest następujący na wstępie każdy element $p \in Z$ reprezentuje osobną klasę. W każdym kolejnym kroku następuje łączenie dwóch najbliższych punktów, w wyniku czego powstaje klastry C do którego należą wszystkie elementy zbioru Z.

- rozdzielającą - polega na dzieleniu grup w punkty;
  Ta metoda jest przeciwieństwem metody aglomeracyjnej, gdyż na wstępie wszystkie elementy zbioru Z należą do jednego klastra. W każdym następnym kroku dochodzi do podziału na mniejsze skupienia w zależności od wartości funkcji oceny odległości.

Grupowanie oparte na gęstościach to metoda w której o podziele zbioru Z na grupy decyduje gęstość położenia elementów należących do zbioru.

Grupowanie znalazło zastosowanie w bankowości, medycynie, przetwarzaniu tekstów. Przykładem zastosowania grupowania może być grupowanie przychodzących wiadomości e-mailowych, gdzie zbiór wiadomości interpretowany jest jako zbiór punktów w przestrzeni wielowymiarowej. Jeden pojedynczy wymiar odpowiada jednemu słowu z określonego słownika. W przestrzeni wielowymiarowej współrzędne wiadomości są zdefiniowane ze względu na częstotliwość występowania słów ze słownika. Klastry odpowiadają grupom dokumentów o tej samej tematyce.

Proces grupowania nie jest pojedynczym procesem, lecz składa się z kilku kroków. Wyróżniamy 4 podstawowe kroki procesu grupowania:

1. wybór reprezentacji obiektów
2. wybór miary podobieństwa pomiędzy obiektami
3. tworzenie klastrów
4. znajdowanie charakterystyki powstałych klastrów

Wybór reprezentacji obiektów, który jest pierwszym krokiem (1), odpowiada za selekcję cech opisujących obiekty, czyli wybór istotnych cech z punktu widzenia procesu grupowania dla użytkownika. Wybranie odpowiednich cech dla danego problemu grupowania. W kroku (2) dokonujemy wyboru miary podobieństwa, dzięki czemu możemy dokonać najlepszego wyboru miary określającej prawdopodobieństwo w zależności od dziedziny zastosowań i grupowanych typów danych. W kolejnym kroku (3) dokonujemy grupowania danych na klastry poprzez wykorzystanie wybranego algorytmu grupowania obiektów, a w kroku (4) ostatnim znajdujemy zwięzłe i czytelne dla użytkownika opisy klastrów.

1.2. Istniejące algorytmy

W tym rozdziale zostaną omówione istniejące algorytmy wykorzystywane do grupowania danych. Przyjrzymy się prostemu algorytmowi K-średnich, a także bardziej złożonym algorytmom, które zostały dostosowane go grupowania.

1.2.1. Algorytmy proste

Istnieje wiele algorytmów grupowania danych, które są wszystkim dobrze znane. Najpopularniejszym algorytmem grupującym jest K-średnich (K-means), który jest przedstawicielem grupowania optymalizacyjno - iteracyjnego.

Algorytm K-średnich to bardo popularny i prosty algorytm grupowania danych, nazywany również algorytmem klastrowy lub LVB (od nazwisk twórców Linde, Buzo i Graya). Pozwala na pogrupowanie zbioru danych Z na N_C klastrów. Działanie algorytmu jest bardzo proste.

Wykorzystane zostaną następujące oznaczenia:
N_d - ilość parametrów (wymiar danych),
Z - zbiór danych do grupowania,
N_C - ilośc klastrów na jakie dzielimy dane,
z_p - p-ty wektor danych ($|Z| = p$; $Z = z_0, ..., z_p$),
m_j - wektor środkowy w klastrze j,
n_j - ilość wektorów danych należących do klastra j,
C_j - zbiór wektorów danych należących do klastra j.

Znając poszczególne oznaczenia przejdźmy do analizy działania algorytmu:

1. losowo wybierz wektory środków klastrów
2. powtarzaj
    (a) oblicz odległości pomiędzy wektorami środkowymi klastrów, a poszczególnymi wektorami danych (data vector)

$$d(z_p, m_j) = \sqrt{\sum_{k=1}^{N_d} (z_{pk} - m_{jk})^2}$$

(b) przelicz wektor środka klastra

$$m_j = \frac{1}{n_j} \sum_{\forall z_p \in C_j} z_p$$

spełniony warunek stopu

Algorytm kończy swoje działanie jeśli spełniony jest jedne z warunków: algorytm wykonał się określoną na wstępie maksymalną liczbę iteracji, występują minimalne zmiany w wektorach środkowych (cluster centroid), lub nie zmienia się ilość członków w klastrach.

Algorytm wykonuje zestaw operacji aż osiągnie jeden z trzech warunków stopu. Wykonywane operacje to obliczanie odległości pomiędzy wektorami środkowymi klastra a danymi do grupowania. Po każdym przeliczeniu odległości dla wektora danych przypisywany jest on do klastra na podstawie najkrótszej obliczonej odległości. Po przeliczeniu członków grup przeliczane jest położenie środka klastra dla każdego klastra. W wyniku działania algorytmu k-średnich otrzymujemy wektory środków klastrów, a także dane pogrupowane w zbioru należące do poszczególnych klastrów.

1.2.2. Algorytmy wykorzystujące inteligencje stadną

Istnieje wiele algorytmów, które zostały wykorzystane do procesu grupowania na podstawie zachowań stadnych np. mrówek, pszczół, ptaków. Możemy wyróżnić algorytm mrowiskowy lub algorytm sztucznej koloni pszczół (Artificial Bee Colony w skrócie ABC), a także algorytm PSO.

Algorytm mrówkowy

Algorytm sztucznej kolonii pszczół inaczej algorytm ABC (Artificial Bee Colony). Popularny algorytm stworzony w 2005 przez Dervis Karaboga. Inspiracją do jego stworzenia było zachowanie kolonii pszczół poszukujących miodu. Algorytm posiada odwzorowanie 3 grup pszczół : pracownic, widzów i zwiadowców. Pszczoły wybierają miejsce zbierania nektaru na podstawie swojego doświadczenia i swoich sąsiadów. W poszukiwaniu nowego miejsca wysyłane są osobniki, które zbierają niezbędne informacje nie mając doświadczenia. Jeżeli znalezione miejsce jest lepsze wysyłana jest informacja o zmianie miejsca zbierania nektaru, a poprzednia lokalizacja jest zapominana.

Algorytm optymalizacji stadnej cząsteczki populanie nazywaney algorytmem PSO. Jego wykorzystanie do grupowania zostanie opisane i zbadane w kolejnych rozdzialach pracy

2. Optymalizacja stadna cząsteczki

Optymalizacja stadna czasteczki (Particle Swarm Optimization) w skrócie PSO to jeden z algorytmów ewolucyjnych, oparty na populacji osobników reprezentujących osobne rozwiązania, należący do szerokiej kategorii algorytmów metod inteligencji stadnej. Został stworzony w oparciu o zaobserwowane społeczne zachowania ptaków.

Twórcami algorytmu optymalizacji stadnej cząsteczki byli R. C. Eberhart i J. Kennedy w 1995 roku. Do inspiracji zachowaniami natury przyczyniła się obserwacja zachować ptaków fruwających w kluczach. Jak zauważono ptaki otrzymują pewną bezpieczną odległość od siebie, oraz pewną prędkość. Jeśli napotkają pewną przeszkodę są wstanie ją bezpiecznie ominąć. Zachowanie danego ptaka jest zależne od jego sąsiadów lub pewnego "przywódcyśtada, czyli najlepszego osobnika. Zachowanie całego stada wynika z zachowania wszystkich ptaków, a nie pojedynczej jednostki, to zachowanie nazywamy zjawiskiem emigracji.

Elementem wyróżniającym PSO pośród innych rozwiązań jest wektor prędkości cząsteczki (ptaka), który odzwierciedla zachowania naturalne, gdzie zmiana prędkości danego osobnika następuje dynamicznie w zależności od najlepiej przystosowanego osobnika. Wektor prędkości zapewnia nam ciągłe poruszanie się w lepszym kierunku.

PSO znajduje swoje zastosowanie do rozwiązywania problemów optymalizacji globalnej, nadaje się do optymalizacji funkcji, a także wykorzystywany jest w procesach grupowania. Na przestrzeni lat prowadzono wiele badań na algorytmem, w wyniku czego powstalo wiele modyfikacji pierszej wersji.

2.1. Opis algorytmu PSO

Jak powyżej opisano optymalizacja stadna cząsteczki, odzwierciedla zachowania ptaków. Stworzony algorytm opisany w [3] zakłada, że populacja początkowa jest populacją losową, gdzie każdy wylosowany element to cząsteczka mieszcząca się w przedziale [min, max]. Następnie ustalana jest najlepsza pozycja lokalna dla każdej cząsteczki, są to najbardziej czasochłonne obliczenia. Konieczne jest również usta-

lenie najlepszej pozycji wśród wszystkich cząsteczek oraz jej zapamiętanie. Każda zmiana położenia cząsteczki związana jest z obliczeniem prędkości. Odbywa się za za pomocą wzoru:

$$\vec{v_{i+1}} = \vec{v_i} + c_1 r_1(\vec{p_l} - \vec{x_i}) + c_2 r_2(\vec{g_g} - \vec{x_i}) \tag{6}$$

gdzie,
$\vec{v_i}$ - wektor prędkości,
c_1, c_2 - stałe przyśpieszenia (dodatnie),
r_1, r_2 - losowe liczy z przedziału [0,1],
$\vec{p_l}$ - wektor najlepszego położenia czasteczki,
$\vec{g_g}$ - wektor najlepszego położenia wśród cząsteczek.

W wzorze na prędkość możemy wyznaczyć dwa współczynniki wpływające na prędkość cząsteczki:

- kognitywny - $c_1 r_1(\vec{p_l} - \vec{x_i})$ - określa stopień wpływu cząsteczki poruszającej się w kierunku swojej najlepszej pozycji
- społeczny - $c_2 r_2(\vec{g_g} - \vec{x_i})$ - określa stopień wpływu cząsteczki poruszającej się w kierunku pozycji globalnie najlepszej.

Ważnym elementem algorytmu PSO jest również $\vec{v_{max}}$, czyli maksymalna prędkość cząsteczki, której cząsteczka nie może przekroczyć. Po obliczeniu prędkości możliwa jest zmiana położenia cząsteczki obliczana zgodzie ze wzorem

$$\vec{x_{i+1}} = \vec{x_i} + \vec{v_{i+1}} \tag{7}$$

Algorytm PSO możemy podzielć na dwa rodzaje ze względu na wpływ sąsiadujących cząsteczek na każdą z cząsteczek:

- gbest -
- lbet -

Pseudokod algorytmu:

2.2. Modyfikacje PSO

Algorytm optymalizacji stadnej cząsteczki w wyniku rozwoju doczekała się kilku modyfikacji, poprzez dodanie pewnych współczynniku polepszających jego działania. Popularne modyfikacje to:

PSO z wagą iteracji

Modyfikacja dodane przez Y. Shi i R. C. Eberharta, poprzez dodanie parametru nazwanego wagą iteracji.

$$\vec{v_{i+1}} = \omega + \vec{v_i} + c_1 r_1(\vec{p_l} - \vec{x_i}) + c_2 r_2(\vec{g_g} - \vec{x_i}), \tag{8}$$

gdzie,
ω - waga iteracji, która powinna malec w czasie z wartości 0,9 do 0,4.

Ustalono również, ze r_1 jak i r_2, powinny być losowana z przedziału [0,2], a nie jak wstępnie ustalono [0,1]. Dzięki wprowadzeniu wagi iteracji możemy zaobserwować jej wpływ na każdą cząsteczkę, podczas zamiany położenia.

Pamiętano również, że wprowadzenie nowego parametru może wpłynąć na działania algorytmu, więc zaproponowano by ω obliczane było ze wzoru

$$\omega = \frac{c_1 + c_2}{2} - 1 \tag{9}$$

Zaproponowana relacja została określona w oparciu o stałe przyśpieszenia.

PSO z wagą ścisku

Modyfikacja zaproponowana przez M. Clerca i J. Kennedy'ego, gdyż obserwacja PSO pokazała, że cząsteczki w pewnym momencie eksplodują i stosowanie tylko v_{max} nie daje oczekiwanych rezultatów. Zaproponowano dodanie współczynnika ścisku występującego we wzorze na prędkość

$$\vec{v_{i+1}} = \chi(\vec{v_i} + c_1 r_1(\vec{p_l} - \vec{x_i}) + c_2 r_2(\vec{g_g} - \vec{x_i})) \tag{10}$$

gdzie, χ - współczynnik ścisku obliczany jest ze wzoru $\chi = \frac{2\kappa}{\left|2-\gamma-\sqrt{\gamma^2-4\gamma}\right|}$, gdzie κ to liczba z przedziału [0,1], a $\gamma = c_1 + c_2$, przy założeniu, że $\gamma > 4.1$

Zaproponowana waga ścisku wpływa nie wszystkie składniki przy obliczaniu prędkości cząsteczki, więc autor stwierdził, że używanie v_{max} jest zbędne.

PSO z selekcją

Modyfikacja zaproponowana przez P. J. Angeline poprzez dodanie selekcji. zaproponowana selekcja mam za zadanie zmniejszyć różnorodność populacji poprzez poddanie każdej cząsteczki ocenie porównania obliczonej wartości przystosowania z wartością losowo wybranej cząsteczki z pewnej grupy. Podczas porównywania zliczane są punktu, które otrzymuje cząsteczka mająca większą wartość. Po wyznaczeniu osobniki są sortowane malejąco względem otrzymanych punktów i dola część osobników zamieniana jest górną.

To przekształcenie sprawdza się w funkcjach unimodalnych, ale nie daje lepszych wyników w rozwiązywaniu problemów z dużą liczbą optimów lokalnych.

PSO w pełni informowalne

Modyfikacja stworzona przez R. Mendes i J. Kennedy nazwana FIPS (ang. Fully Informed PSO). Powstała w wyniku obserwacji, że tak naprawdę nie kierujemy się pod wpływem jednego osobnika, a statyczną obserwacją sąsiadów. Jest to całkowite inne podejście niż standardowe PSO, gdyż został zaproponowany nowy wzór na prędkość

$$\vec{v_{i+1}} = \chi(\vec{v_i} + c_m(\vec{p_m} - \vec{x_m})) \tag{11}$$

gdzie:

$c_m = c_1 + c_2$,

$\vec{p_m} = \frac{c_1 * \vec{p_i} + c_2 * \vec{p_g}}{c_1 + c_2}$

Wzór na aktualne położenie cząsteczki to

$$\vec{x_{i+1}} = \vec{x_i} + \vec{v_{i+1}} \tag{12}$$

gdzie:

$\vec{x_{i+1}}$ - nowa pozycja cząsteczki w chwili $t + 1$

$\vec{x_i}$ - aktualna pozycja cząsteczki

$\vec{v_{i+1}}$ - prędkość cząsteczki w chwili $t + 1$

Dzięki temu mamy pewność, że na poszczególne cząsteczki mają wpływ wyniki sąsiadów.

2.3. Topologie komunikacyjne

Topologie komunikacyjne to pojęcie pozwalające określić, jak duży zakres cząsteczek będzie mieć wpływ na zmiany ruchu aktualnie rozpatrywanej cząsteczki.

Wybór odpowiedniej topologii może mieć znaczący wpływ na działanie algorytmu. Wyróżniamy następujące topologie :

- topologia lbest
- topologia gbest
- topologia gwiazdy
- topologia pierścienia
- topologia Von Neumana
- model wyspowy

Topologia lbest to topologia, gdzie wpływ na ruch analizowanej cząsteczki ma jej najlepszy dotychczasowy wynik, a także najlepszy wynik spośród osobników w najbliższym otoczeniu. W topologii gbest wpływ na cząsteczkę ma jej najlepszy wynik, a także najlepszy wynik spośród wszystkich osobników. W topologii gwiazdy zostaje wybrana jedna cząsteczka, która staje są cząsteczką - hub-em do której podłączone są wszystkie cząsteczki populacji. Topologia pierścienia zorganizowana jest w pierścień, a sąsiedztwo danej cząsteczki określone jest przez liczbę prawych i lewych jej sąsiadów. Cząsteczki połączone ze sobą za pomocą sieci, gdzie każda cząsteczka jest połączona możemy zobaczyć w topologii Von Neumana. Natomiast model wyspowy to rozwiązanie umożliwiające grupowanie osobników w wyspy, na których można stosować wcześniej omówione topologie.

3. PSO w grupowaniu danych

Celem grupowanie jak wspomniano w rozdziale 1.1 jest znalezienie **Ocena jakości grupowania**

4. Porównanie klasycznego PSO do algorytmu PSO w grupowaniu danych

5. Implementacja

6. Badania

Literatura

[1] http://edu.pjwstk.edu.pl/wyklady/adn/scb/wyklad13/w13.htm

[2] artykuł : Data Clustering using Particle Swarm Optimization

[3] Andries P. Engelbrecht, Computational Intelligence An Introduction, wyd. 2, Anglia, WILEY, ISBN 978-0-470-03561-0 s. 285 - 359.