# Optymalizacja roju cząstek dla wykrywania wartości odstających

**Dokument konferencyjny** - lipiec 2010 r.

DOI: 10.1145/1830483.1830498 - Źródło: DBLP



**3 autorów:**

Ammar Mohemmed
Politechnika w Auckland
**31** PUBLIKACJI **1 071** CYTOWAŃ

ZOBACZ PROFIL

Mengjie Zhang
Uniwersytet Wiktorii w Wellington
**881** PUBLIKACJI **19 636** CYTOWAŃ

ZOBACZ PROFIL

Will Neil Browne
Queensland University of Technology
**197** PUBLIKACJI 4 **784** CYTOWAŃ

ZOBACZ PROFIL

Cała zawartość tej strony została dodana przez Will Neil Browne 01 marca 2014.

Użytkownik zażądał ulepszenia pobranego pliku.

# Optymalizacja roju cząstek dla wykrywania wartości odstających

Ammar W Mohemmed, Mengjie Zhang, Will Browne
School of Engineering and Computer Science, Victoria University of Wellington
PO Box 600, Wellington, Nowa Zelandia

Raport techniczny: ECSTR10-07

## ABSTRAKT

Wykrywanie wartości odstających jest ważnym problemem, ponieważ bazowe punkty danych często zawierają kluczowe informacje, ale identyfikacja takich punktów wiąże się z wieloma trudnościami, np. zaszumionymi danymi, nieprecyzyjnymi granicami i brakiem przykładów szkoleniowych. W tym nowatorskim podejściu problem wykrywania wartości odstających jest przekształcany w problem optymalizacji. Następnie można zastosować podejście oparte na optymalizacji rojem cząstek (PSO) do wykrywania wartości odstających, co rozszerza zakres PSO i umożliwia nowe spojrzenie na wykrywanie wartości odstających. Mianowicie, PSO jest używane do automatycznej optymalizacji kluczowych miar odległości zamiast ręcznego ustawiania parametrów odległości metodą prób i błędów, co jest nieefektywne i często nieskuteczne. Nowatorskie podejście PSO zostało zbadane i porównane z powszechnie stosowaną metodą wykrywania, Local Outlier Factor (LOF), na pięciu rzeczywistych zestawach danych. Wyniki pokazują, że nowa metoda PSO znacznie przewyższa metody LOF pod względem prawidłowego wykrywania wartości odstających w większości zestawów danych oraz że nowa metoda PSO jest bardziej wydajna niż metoda LOF w testowanych zestawach danych.

## Kategorie i deskryptory tematów

I.5 [Rozpoznawanie wzorców]; I.2 [Sztuczna inteligencja]

## Warunki ogólne

Algorytmy, projektowanie

## Słowa kluczowe

Wykrywanie wartości odstających, optymalizacja rojem cząstek

## 1. WPROWADZENIE

Dokładna definicja wartości odstającej często zależy od kontekstu domeny aplikacji i zastosowanej metody wykrywania. Jednak definicja Hawkinsa jest uważana za wystarczająco ogólną, aby poradzić sobie z różnymi aplikacjami i metodami: *Obserwacja odstająca to obserwacja, która odbiega tak bardzo od innych obserwacji (uważanych za normalne), że wzbudza podejrzenie, że była to obserwacja normalna.*

*generowane przez inny mechanizm* [7]. Znaczenie wykrywania wartości odstających wynika z faktu, że wartości odstające w danych ujawniają ukryte informacje, które czasami wymagają szybkiego działania, aby uniknąć przyszłych szkód lub szkód. Na przykład w finansach wykrywanie wartości odstających, takich jak oszustwa związane z kartami kredytowymi, może uruchomić działania zapobiegające większej utracie pieniędzy przez klientów i banki. W przypadku bezpieczeństwa sieci, jak najwcześniejsze wykrycie podejrzanych zachowań związanych z włamaniami może zapobiec większym uszkodzeniom sieci i jej komponentów. Wczesne diagnozowanie usterek w maszynach (silnikach, generatorach, promach kosmicznych itp.) może uratować ludzkie życie i zapobiec katastrofalnym uszkodzeniom, a także wielu innym zastosowaniom.

Problem wykrywania wartości odstających, w swojej najbardziej ogólnej formie, jest trudny do rozwiązania ze względu na szereg wyzwań [5]. Granica między zachowaniem normalnym a odstającym często nie jest precyzyjna. W wielu domenach problemowych normalne zachowania ewoluują, a dokładne pojęcie wartości odstającej różni się w zależności od zadania. Często trudno jest uzyskać wystarczającą ilość danych odstających do szkolenia i/lub walidacji. Dane często zawierają szum, który sprawia, że normalne obserwacje stają się podobne do rzeczywistych wartości odstających i odwrotnie.

Zaproponowano wiele technik wykrywania out- lierów dla różnych zastosowań. Techniki te można podzielić na kilka podejść [5, 9, 3]: *podejście statystyczne, podejście oparte na grupowaniu* i *podejście oparte na odległości*. W podejściu opartym na odległości obliczana jest prosta odległość lub miara podobieństwa między każdymi dwoma instancjami/punktami w zbiorze danych, a punkty, których odległości są dłuższe niż określony promień (próg), są uważane za wartości odstające. W porównaniu z podejściem opartym na grupowaniu, podejście oparte na odległości jest znacznie prostsze w użyciu. W przeciwieństwie do metod statystycznych, metody oparte na odległościach nie przyjmują żadnych wcześniejszych założeń dotyczących modelu dystrybucji danych i są bardziej odpowiednie dla wielowymiarowych zbiorów danych. Ze względu na te zalety, podejście oparte na odległości jest szeroko stosowane w wykrywaniu wartości odstających.

W podejściu opartym na odległości istnieją dwie podstawowe metody, na których opiera się wiele późniejszych technik [19, 20, 2, 17]. Pierwsza z nich opiera się na pracy Knorra i in.[11]. W tej pracy punkt danych jest definiowany jako wartość odstająca oparta na odległości DB($\beta$,$r$), jeśli co najmniej ułamek $1 \boxtimes \beta$ instancji w zbiorze danych znajduje się dalej niż $r$ od niego, gdzie $\beta$ jest określane przez użytkownika na podstawie rzeczywistej sytuacji, a $r$ jest promieniem odległości działającym jako próg wartości odstającej. Podczas gdy $\beta$ jest stosunkowo łatwe do określenia, ponieważ wartości odstające mają zwykle małe sąsiedztwo, wartość $r$ jest zwykle bardzo trudna do oszacowania i zazwyczaj wymaga prób i błędów poprzez ręczne wyszukiwanie empiryczne [19].

Druga podstawowa metoda stanowi rozwinięcie pracy Breunig et al. [4]. Praca ta wykorzystuje lokalne czynniki odstające (LOF) zamiast globalnych odległości [11]. Punkt danych otrzymuje ocenę wartości odstającej na podstawie jego względnej gęstości w odniesieniu do najbliższych punktów sąsiednich. Metoda ta może wykrywać wartości odstające w zbiorach danych, które mają regiony o różnej gęstości, co nie może być łatwo obsługiwane przez algorytm Knorra [11]. Metoda ta wymaga jednak określenia liczby punktów sąsiedztwa (*MinPtn*) *a priori,* co zazwyczaj wymaga ręcznego wykonania oraz prób i błędów.

Inną potencjalną wadą metod opartych na odległości jest koszt obliczeniowy. W przypadku stosunkowo małych zbiorów danych nie stanowi to problemu. Jednak w przypadku większych zbiorów danych metody te zazwyczaj wymagają dużego wysiłku obliczeniowego, ponieważ obliczenie odległości między dużą liczbą instancji/punktów danych jest kosztowne [9].

W typowych metodach wykrywania wartości odstających opartych na odległości, głównym zadaniem jest znalezienie dobrych wartości ważnych parametrów, takich jak $\beta$, $r$ i *MinPtn* opisanych wcześniej. Zadanie to można naturalnie przełożyć na problem optymalizacyjny, który można rozwiązać za pomocą niektórych paradygmatów obliczeń ewolucyjnych, takich jak algorytmy genetyczne i optymalizacja rojem cząstek. W ostatnich latach wykonano tylko niewielką ilość prac z zastosowaniem technik obliczeń ewolucyjnych do procesu wykrywania wartości odstających, ale były one wykorzystywane głównie do redukcji wymiaru i selekcji cech [15, 1, 21]. Niniejszy artykuł ma na celu przekształcenie problemu wykrywania wartości odstających w problem optymalizacji i opracowanie podejścia optymalizacji roju cząstek (PSO) do wykrywania wartości odstających przy użyciu miar opartych na odległości. Zamiast używać ręcznego procesu prób i błędów, podejście to automatycznie ewoluuje dobre wartości dla ważnych parametrów. W celu zbadania tego podejścia zostanie ono przeanalizowane i porównane z powszechnie stosowaną metodą wykrywania wartości odstających opartą na odległości (LOF) na sekwencji wykrywania wartości odstających.
problemy. W szczególności zbadamy:

- W jaki sposób cząstki w populacji mogą być kodowane dla zadania wykrywania wartości odstających;
- Jak poszczególne cząstki są oceniane podczas procesu ewolucji;
- Czy to podejście przewyższa metodę LOF w sekwencji zadań wykrywania wartości odstających; i
- Czy to podejście jest bardziej wydajne niż metoda LOF, szczególnie w przypadku dużych zbiorów danych.

# 2. TŁO

## 2.1 Problem wykrywania wartości odstających

Wykrywanie wartości odstających nie ma jak dotąd uzgodnionej definicji. W tym artykule używamy definicji Hawkinsa, która jest wystarczająco ogólna, aby poradzić sobie z różnymi zastosowaniami i metodami [7], jak opisano we wstępie.

Problem wykrywania wartości odstających jest podobny, ale różni się od klasyfikacji binarnej. W klasyfikacji binarnej klasyfikator jest trenowany na wystarczającej liczbie negatywnych i pozytywnych przykładów zbioru danych, aby uchwycić jego charakterystykę, a następnie używany do klasyfikowania niewidocznych obiektów. Wydajność klasyfikatora jest zwykle oceniana na podstawie jego dokładności klasyfikacji, poziomu błędu całego zbioru danych lub frakcji wyników prawdziwie pozytywnych i fałszywie pozytywnych dla określonej klasy. W przypadku wykrywania wartości odstających zadanie jest jednak nieco bardziej rozmyte. Celem jest zidentyfikowanie obiektów, które *różnią* się od reszty zbioru danych.

dane, które są uważane za normalne lub zwykłe. Nie jest jednak jasne, w jaki sposób i w jakim stopniu obiekty powinny się różnić, aby uznać je za wartości odstające. Powszechną praktyką wykrywania wartości odstających jest nadanie punktom danych wyniku odstającego, a następnie posortowanie ich w celu wyodrębnienia potencjalnych wartości odstających na szczycie listy. Różni się to od typowych zadań klasyfikacji binarnej, w których jedna instancja danych jest uważana za poprawną lub niepoprawną w stosunku do określonej klasy.

Jednak ze względu na podobieństwo między wykrywaniem wartości odstających a klasyfikacją binarną, wiele "wzorcowych zestawów danych do wykrywania wartości odstających" opiera się na niektórych zadaniach klasyfikacji binarnej z wysoce niezrównoważonymi danymi dla dwóch klas jako poligonem testowym [8]. W niniejszym artykule zostaną również wykorzystane niektóre z tych zestawów danych, które zostaną opisane w sekcji 4.

## 2.2 Powiązane prace

Zaproponowano różne techniki wykrywania wartości odstających. Początkowo techniki te były zdominowane przez metody statystyczne [7, 3]. Metody te zakładają określony model rozkładu i są w większości jednowariantowe, tj. badają pojedynczy atrybut w celu wykrycia wartości odstającej. Metody te nie nadają się do zastosowań wielowymiarowych [11]. Metody klastrowania do wykrywania wartości odstających opierają się na założeniu, że większość normalnych danych należy do jednego lub więcej klastrów, podczas gdy wartości odstające albo stanowią bardzo mały klaster, albo są daleko od głównych utworzonych klastrów. Jednak podejścia oparte na grupowaniu często znajdują wartości odstające jako produkt uboczny głównego procesu, więc większość z nich nie jest zoptymalizowana pod kątem wykrywania wartości odstających [4].

Podejście oparte na odległości zaproponowane przez Knorra [11] jest szeroko cytowane w literaturze dotyczącej wykrywania wartości odstających. Chociaż opiera się ono na prostej koncepcji, nie zakłada konkretnego modelu rozkładu, a także może obsługiwać dane o dużej wymiarowości. Biorąc pod uwagę zbiór danych, punkt jest uważany za odstający, jeśli ułamek 1 ⊠ $\beta$ punktów danych znajduje się dalej niż $r$ (próg odległości) od tego punktu. Dwa parametry $r$ i $\beta$ są określane przez użytkownika. Określenie $r$ nie jest proste, więc początkowo konieczne jest zastosowanie metody prób i błędów w celu określenia odpowiedniej wartości. Innym problemem związanym z tym podejściem jest to, że nie zapewnia ono sposobu na uszeregowanie wartości odstających, ponieważ albo kategoryzuje punkt jako odstający, albo nie.

Ramaswamy i in. proponują algorytm [19], który pomija wymóg określenia $r$ przez użytkownika powyżej. Punkty danych otrzymują wynik odstający na podstawie odległości od *MinPtn* ⊠ najbliższych punktów. Algorytm ten najpierw dzieli wejściowy zbiór danych na rozłączne podzbiory za pomocą klastrowania, a następnie przycina te partycje określone na podstawie tego, czy zawierają wartości odstające, pozostawiając partycje kandydujące, które mogą zawierać wartości odstające. Ten krok ma na celu przyspieszenie obliczeń, szczególnie w przypadku bardzo dużych zbiorów danych, ponieważ wiele punktów zostanie wyeliminowanych, więc nie ma potrzeby znajdowania *MinPtn-tego* najbliższego punktu dla tych wyeliminowanych punktów danych. Wreszcie, wartości odstające są obliczane spośród punktów w partycjach kandydujących. Metody oparte na DB [11] mogą nie wykrywać wartości odstających w zbiorach danych składających się z różnych regionów gęstości, ponieważ uwzględniają punkty danych globalnie.

Breunig et al. [4] proponują lokalny algorytm wykrywania. Punkty danych są oceniane za pomocą współczynnika LOF (Local Outlier Factor), który reprezentuje stopień odstających punktów w zależności od ich lokalnego sąsiedztwa. Dla dowolnego punktu danych wynik LOF jest równy stosunkowi średniej lokalnej gęstości *MinPtn* najbliższych sąsiadów punktu i lokalnej gęstości samego punktu danych. Lokalna gęstość

Gęstość punktu, nazywana *gęstością osiągalności*, jest definiowana jako odległość osiągalności. Odległość osiągalności jest obliczana na podstawie odległości do *MinPtn-tego* najbliższego sąsiedztwa. W przypadku normalnego punktu danych leżącego w gęstym regionie, jego lokalna gęstość osiągalności będzie podobna do gęstości jego sąsiadów i będzie miała niski LOF, podczas gdy w przypadku wartości odstających jego lokalna gęstość będzie niższa niż gęstość jego najbliższych sąsiadów, a zatem uzyska wyższy wynik LOF. Wybór *MinPtn jest jednak* nietrywialny, a koszt obliczeń jest bezpośrednio związany z wartością *MinPtn*.

Idea LOF została rozszerzona i ulepszona na różne sposoby. Na przykład schemat Connectivity-based Outlier Factor (COF) [20] rozszerza algorytm LOF do wykrywania wartości odstających we wzorcach danych, które są trudne do roz- krycia za pomocą LOF. LSC-Mine [2] upraszcza obliczanie lokalnej gęstości osiągalności wraz z przycinaniem, co prowadzi do szybszych obliczeń.

Podobna technika o nazwie LOCI (Local Correlation Inte- gral) została przedstawiona w [17]. LOCI rozwiązuje trudność wyboru *MinPtn* w technice LOF, stosując inną definicję sąsiedztwa. Dla każdego punktu danych badane jest sąsiedztwo w obrębie różnych wartości *r*. Punkt jest oznaczany jako odstający, jeśli parametr zwany współczynnikiem odchylenia wieloziarnistości (MDEF) trzykrotnie odbiega od odchylenia standardowego MDEF w sąsiedztwie. MDEF w promieniu *r* dla punktu to względne odchylenie jego lokalnej gęstości sąsiedztwa od średniej lokalnej gęstości sąsiedztwa w jego sąsiedztwie *r*. Zatem obiekt, którego gęstość sąsiedztwa odpowiada średniej lokalnej gęstości sąsiedztwa, będzie miał MDEF równy 0. Natomiast wartości odstające będą miały MDEF dalekie od 0. Jednak koszt działania algorytmu jest wysoki ze względu na potrzebę obliczania wartości statystycznych, w tym odchylenia standardowego. Niedawno Zhang et al. [22] zaproponowali Local Distance-based Outlier Factor (LDOF) do pomiaru odstających obiektów w rozproszonych zbiorach danych. LDOF wykorzystuje względną lokalizację obiektu do jego sąsiadów, aby określić stopień, w jakim które obiekt odchyla od swojego sąsiedztwa. Kriegel et al. [12] proponuje wykorzystanie wariancji kątów między wektorami różnicowymi punktu do innych punktów w celu identyfikacji wartości odstających. Intuicja stojąca za tą metodą polega na tym, że dla punktów w klastrze kąty między wektorami różnic do par innych punktów różnią się znacznie, w przeciwieństwie do wartości odstających, które będą miały małą wariancję kątów. Aby zredukować oceny odległości parami w technikach opartych na odległości i gęstości, metoda zaproponowana przez Yaling [18] definiuje stały zestaw punktów odniesienia do rankingu punktów danych i wykrywania wartości odstających.

Algorytmy ewolucyjne, takie jak algorytmy genetyczne (GA), zostały wykorzystane do wykrywania wartości odstających. Kelly et al. [6] wykorzystali GA do przeszukiwania danych regresji w celu znalezienia podzbioru punktów o najwyższym stopniu dopasowania/odstających. Funkcja fitness jest testem diagnostycznym dla wartości odstających.

Metody oparte na przykładach pozwalają użytkownikom na dostarczenie technice pewnych odstających przykładów w celu znalezienia większej liczby obiektów, które mają podobne cechy odstające do tych przykładów. Yuan et al. [15] wykorzystują GA do znalezienia niskowymiarowej podprzestrzeni, w której dane przykłady użytkownika są izolowane bardziej znacząco niż w innych podprzestrzeniach, a następnie wykrywają wartości odstające DB($\beta$,$r$) w tej podprzestrzeni. Wadą tego algorytmu jest to, że nadal istnieje potrzeba interwencji użytkownika w celu wybrania parametrów do decydowania o wartościach odstających.

aby znaleźć niskowymiarowe kostki o najniższej rzadkości danych. Ten sam algorytm został zaimplementowany przez Dongyi et al. [21], ale przy użyciu PSO. Wadą tego algorytmu jest założenie, że punkty danych są równomiernie rozłożone, a współczynnik rzadkości opiera się na rozkładzie normalnym.

## 2.3 Optymalizacja rojem cząstek

Optymalizacja rojem cząstek (PSO) jest stochastycznym narzędziem optymalizacyjnym opartym na populacji, inspirowanym zachowaniem społecznym stad ptaków (i ławic ryb itp.), opracowanym przez Kennedy'ego i Eberharta w 1995 roku [10]. PSO składa się z populacji cząstek, które szukają rozwiązania w domenie wyszukiwania. Poszukiwanie optymalnej pozycji (rozwiązania) odbywa się poprzez aktualizację prędkości $V_i$ i pozycji Xi cząstki *i* zgodnie z następującymi dwoma równaniami:

$$V_i = V_i + \varphi 1 c1(X_i^{best} \boxtimes X_i) + \varphi 2 c2(X_i^{gbest} \boxtimes X_i) \quad (1)$$

$$X_i = X_i + V_i \quad (2)$$

gdzie $\varphi 1$ i $\varphi 2$ są dodatnimi stałymi, zwanymi *współczynnikami przyspieszenia*, $c1$ i $c2$ są dwiema niezależnie wygenerowanymi liczbami porządkowymi z zakresu [0, 1], $X^{best}$ jest najlepszą pozycją *i-tej* cząstki, a $X^{gbest}$ jest najlepszą pozycją znalezioną przez sąsiedztwo cząstki *i*. Rój rozpoczyna się od inicjalizacja prędkości i pozycji cząstek losowo z wartościami ograniczonymi przez domenę wyszukiwania. Następnie rój przechodzi przez pętlę, podczas której pozycje i prędkości są aktualizowane zgodnie z powyższymi równaniami. Gdy kryterium zakończenia zostanie spełnione, najlepsza cząstka (z jej pozycją) znaleziona do tej pory jest traktowana jako rozwiązanie problemu. Aby zapobiec eksplozji prędkości, jest ona zwiększana do $V^{max}$.

Alternatywnie, Clerc et al. [16] zaproponowali dodanie współczynnika ograniczenia, aby zapobiec przekroczeniu przez prędkości granic domeny wyszukiwania. W ten sposób równanie aktualizacji prędkości jest modyfikowane w następujący sposób:

$$V_i = \chi \times V_i + \varphi 1 c1(X_i^{best} \boxtimes X_i) + \varphi 2 c2(X_i^{gbest} \boxtimes X_i) \quad (3)$$

Aggrawal et al. [1] definiuje współczynnik rzadkości przy użyciu GA

gdzie $\chi$ jest współczynnikiem zwężenia i zwykle wynosi 0,729.

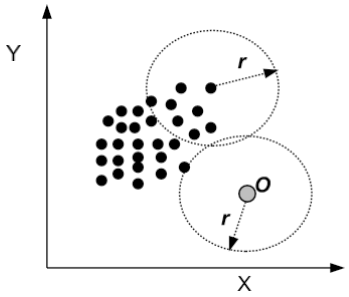
Topologia sąsiedztwa roju określa, w jaki sposób cząstki są ze sobą połączone i wpływa na sposób wymiany informacji między cząstkami. W topologii pierścienia każda cząstka jest połączona z dwoma innymi sąsiadami, tworząc pierścień. W topologii globalnej każda cząstka jest połączona ze wszystkimi innymi cząstkami. Przewaga topologii pierścieniowej nad topologią globalną polega na tym, że transfer informacji między cząstkami jest powolny, co pomaga uniknąć popadnięcia roju w lokalne optimum. Dlatego w tym artykule zastosowano topologię pierścieniową.

## 3. NOWE PODEJŚCIE OPARTE NA PSO

Nasze podejście oparte na PSO wykorzystuje miarę opartą na odległości. Aby zilustrować, które punkty danych można uznać za odstające podczas ewolucji PSO, użyjemy przykładowego zestawu danych pokazanego na rys. 1. W tym przykładzie punkt danych *o* jest bardziej prawdopodobny jako punkt odstający, ponieważ różni się i jest daleko od innych punktów. Widoczna różnica polega na liczbie pobliskich punktów, które znajdują się w określonym promieniu/odległości *r* od niego. Zakładając okrąg o promieniu *r* wyśrodkowany na punkcie danych, punkt *o* ma minimalny współczynnik$\frac{k}{}$ , gdzie *k* jest liczbą punktów zamkniętych w okręgu o promieniu *r*. Innymi słowy, współczynnik$\frac{k}{}$ może być wykorzystany jako miara do wykrywania wartości odstających.

*r*

*r*




Rysunek 1: Przykładowy zestaw danych z punktami odniesienia.

Należy zauważyć, że pomysł ten jest powiązany z definicją wartości odstającej Knorra [11]: punkt danych jest wartością odstającą, jeśli ma ułamek $\beta$ lub mniej z całkowitej liczby punktów w odległości $r$. W rzeczywistości $\beta$ w metodzie Knorra jest funkcją $k$ wspomnianą powyżej, która zmienia się wraz z wartością $r$. Wewnętrzna zależność sprawia, że ręczne ustawienie dobrych wartości dla dwóch parametrów jest jeszcze trudniejsze. W związku z tym, zamiast decydowania przez użytkownika o wartości $\beta$ i $r$, proponujemy algorytm wykorzystujący PSO do automatycznego znajdowania odpowiednich wartości na podstawie danych.

W tym podejściu PSO jest używane do znalezienia punktu, który ma minimum $k_r$ stosunek. Wartość $r$, która skutkuje minimalną wartością $k$ jest używana do obliczenia współczynnika dla innych punktów i odpowiedniego ich uszeregowania w celu zidentyfikowania najlepszych wartości odstających. W prostym przykładzie pokazanym na rysunku 1 wartość $r$ można ustawić na odległość najbliższego sąsiedztwa do $o$, ponieważ wystarczyłoby to do oceny punktów danych i umieszczenia $o$ na szczycie wykrytych wartości odstających. Niestety, większość rzeczywistych zestawów danych zawiera znacznie bardziej złożone rozkłady, które sprawiają, że znalezienie $r$ *jest* bardzo trudnym zadaniem.

Intuicyjnie, algorytm PSO można zaimplementować w celu znalezienia pojedynczego parametru $r$. W takim przypadku cząstka może zakodować tylko jeden parametr, $r$. Aby ocenić dobroć cząstki, dla każdego punktu danych obliczany jest współczynnik $k_r$. Następnie Minimalna wartość współczynnika $k$ wśród punktów danych może być wykorzystana jako funkcja fitness. Punkt danych o najlepszej wartości $r$, który prowadzi do minimalnej wartości współczynnika $k_r$, może zostać wykryty jako najbardziej odstający. Jednak obliczanie współczynnik $k_r$ dla wszystkich punktów danych za każdym razem, gdy było to konieczne do ocena wszystkich cząstek w populacji może być bardzo nieefektywne obliczeniowo. Konieczne było opracowanie lepszego schematu kodowania i odpowiedniej funkcji fitness.

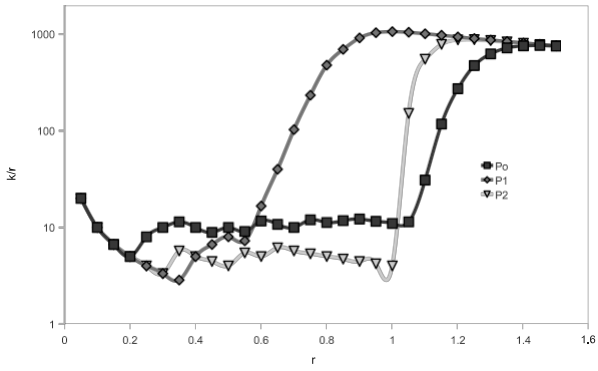
## 3.1 Kodowanie cząsteczek

Aby uzyskać bardziej wydajny schemat kodowania, rozważamy inny sposób dalszego wykorzystania możliwości wyszukiwania PSO. Oprócz parametru promienia $r$, kodowanie cząstki jest teraz rozszerzone o indeks punktu danych. Tak więc cząstka koduje krotkę *ID, r*. Parametr *ID* jest indeksem punktów danych. Jest to wartość całkowita, więc wartości zwracane przez cząstki są zaokrąglane w górę do najbliższej wartości całkowitej. $r$ jest promieniem hipersfery wyśrodkowanej na *ID* punktu danych.

Celem wyszukiwania jest znalezienie punktu danych który ma minimum $k_r$ i jednocześnie znaleźć wartość $r$, która skutkuje minimalizacją tego współczynnika. Uwzględniając

## 3.2 Funkcja fitness

### *3.2.1 Rozważania projektowe*




Rysunek 2: Zmiana $k_r$ z $r$ dla 3 punktów ze zbioru danych drożdży, P0 pochodzi z klasy mniejszościowej, P1 i P2 z klasy większościowej.

Podczas projektowania funkcji fitness należy wziąć pod uwagę szereg kwestii. Przeanalizujemy je na przykładzie

ple. Rys. 2 przedstawia zmianę współczynnika $k_r$ dla trzech punktów z zestawu danych Yeast [14]. Oryginalny zestaw danych Yeast zawiera dziewięć klas. Wśród nich pierwsze trzy klasy mają łącznie 1136 przykładów, a ostatnia klasa składa się tylko z pięciu instancji danych. Aby wykorzystać go do wykrywania wartości odstających, stworzyliśmy nowy zestaw danych oparty na oryginalnym zestawie. Nowy zestaw danych wykorzystuje 1136 punktów danych w pierwszych trzech klasach jako *normalną* grupę, a pięć punktów danych w ostatniej klasie jako wartości odstające, które chcemy wykryć. Wykres współczynnika $k_r$ w odniesieniu do $r$ dla jednego z rozważanych punktów odstających (P0) pokazano na rys. 2. Pozostałe dwa punkty (P1 i P2) należą do grupy normalnej.

Globalna minimalna wartość $k_r$ wynosi $r$ równe 0,35. Wartość ta jest jednak spowodowana punktem P1, który należy do jednej z grup normalnych. Dlatego PSO zwróci P1 (i $r$ = 0,3) jako punkt odstający i będzie w wyższej randze niż P0, co oczywiście nie jest tym, czego chcieliśmy. Minimalna wartość

$r$ dla punktu danych P0 wynosi 0,2, ale P1 i P2 również

ta sama wartość przy $r$ = 0,2 dla , co nie pomaga w sklasyfikowaniu P0 wyżej niż P1 i P2. Co ważne, zbyt mała wartość $r$ nie jest dobra do wykrywania wartości odstających, nawet jeśli wartość $k_r$ jest najmniejsza lub bardzo mała. Z drugiej strony, wartość $r$ nie powinna być tak duża, aby obejmowała wszystkie punkty zestawu danych. W takim przypadku współczynnik $k_r$ będzie taki sam dla wszystkich punktów danych, co również nie pomoże w ich uszeregowaniu. Dlatego $r$ powinno mieć dolną i górną granicę, aby zidentyfikować prawidłowe wartości odstające. W związku z tym, podczas projektowania funkcji fitness, powinniśmy rozważyć tę kwestię oprócz kluczowej miary $k_r$.

### *3.2.2 Nowa funkcja fitness*

Funkcja fitness składa się z trzech następujących elementów:

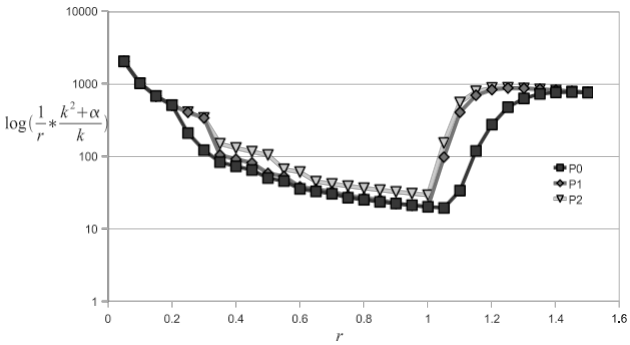
$$\alpha \quad k \quad k \qquad (4)$$

indeks punktu jako kolejna zmienna wyszukiwania oprócz $r$, nie jest konieczne obliczanie współczynnika $k_r$ dla

wszystkich punktów za każdym razem, gdy oceniana jest cząstka. Oczekujemy, że PSO automatycznie znajdzie mniejszą grupę "zoptymalizowanych" punktów danych, które mają lepszy potencjał jako wartości odstające.

gdzie $n$ jest rozmiarem zbioru danych, a $\alpha$ jest stałą. Pierwszy człon $r_{r*k}^{\alpha}$, gdzie $\alpha$ jest stałą ograniczającą dolną granicę $r_{r*k}$. Wartość $r$ powinna być wystarczająco duża, aby uwzględnić niektóre sąsiednie punkty, aby można było obliczyć stosunek $\frac{\alpha}{r*k}$ i wykryć wartości odstające. Drugi termin $\frac{k}{\gamma}$ jest kluczową miarą

do optymalizacji. Te dwa terminy są połączone jako $\frac{1}{r}\left(\frac{k}{r}+\alpha k\right)$ narysowane względem $r$ na Rys.3 dla trzech punktów z Rys.2.
Z tego wykresu widać, że dodając pierwszy do kluczowej miary (drugi termin), minimum (przy $r$ = 1,05) wynika z punktu P0, który jest prawdziwym punktem odstającym. Tak więc na ostatecznej liście rankingowej punkt P0 znajdzie się przed punktami P1 i P2. Pierwszy człon (a zatem stała $\alpha$) wpłynie na minimalną liczbę punktów, które zostaną uwzględnione przez $r$. Ustawiając wartość $\alpha$ do $0{,}05 \times n$, zapewnimy, że $r$ dla wartości odstających będzie wynosić wystarczająco duży, aby uwzględnić niektóre sąsiednie punkty.




Rysunek 3: Zmiana $\frac{1}{r}\left(\frac{k}{n-k}+\alpha k\right)$ z $r$ dla 3 punktów z Rys.2.

Trzeci termin ma na celu ograniczenie górnej wartości granicznej $r$. Należy zauważyć, że ten termin jest skuteczny, gdy $r$ jest duże, więc obejmuje dużą liczbę punktów danych. Na przykład, gdy $r$ jest duże, tak że $k = n$, wartość tego terminu wyniesie $\infty$, więc cząstki zostaną odepchnięte od tych punktów, które najprawdopodobniej nie wskazują na wartości odstające. Dlatego $n$ powinno być zwykle znacznie większe niż $k$, szczególnie w przypadku punktów odstających, w których wartość terminu jest bardzo mała. W związku z tym pierwszy i trzeci człon mają na celu ograniczenie wartości $r$ do zakresu przydatnego do wykrywania wartości odstających. W tym zakresie $r$ nie jest zbyt małe, co mogłoby prowadzić do pominięcia wartości odstających i nie jest zbyt duże, aby uwzględnić zbyt wiele fałszywie dodatnich punktów z grupy normalnej.

Należy pamiętać, że mogliśmy dodać współczynnik działający jako współczynnik wagi do każdego z trzech terminów, aby jeszcze bardziej odzwierciedlić względne znaczenie między kluczową miarą $k/r$ a dwiema granicami $r$. Chociaż może to potencjalnie pomóc w poprawie wydajności procesu ewolucyjnego, jeśli można znaleźć dobre wartości wag, wyszukiwanie tych współczynników będzie wymagało dalszej ręcznej konfiguracji metodą prób i błędów. W rzeczywistości traktowanie ich jako równie ważnych (ustawienie wszystkich na 1,0 jak w równaniu 4) może osiągnąć dobre wyniki, ponieważ PSO może automatycznie zoptymalizować $r$. Dlatego użyjemy równania 4 jako funkcji fitness.

## 3.3 Algorytm outPSO

Łącząc wszystkie aspekty razem, cały nowy algorytm oparty na PSO do wykrywania wartości odstających został przedstawiony w algorytmie 1. Algorytm rozpoczyna się od losowej inicjalizacji cząstek z pierwszym parametrem, $x_1$, inicjowanym w zakresie [0, $n$]. Drugi parametr, $x_2$, jest inicjowany w zakresie [0, *maxdist*], gdzie *maxdist* jest maksymalną możliwą odległością między dowolnymi dwoma punktami. Maksymalne wartości prędkości są ustalane na podstawie wstępnych wyników eksperymentalnych. Algorytm działa do momentu osiągnięcia maksymalnej liczby iteracji.

Algorytm 1 Pseudokod dla wykrywania wartości odstających w oparciu o PSO outPSO

Ustaw $x^{min} = 0$, $x^{max} = n$ i $x^{min} = 0$, $x^{max} = maxdist$
Ustaw $v_1^{min} = -1.0$, $v_1^{max} = +1.0$ i $v_2^{min} = -1.0$, $v_2^{max} = +1.0$
{maxdist to maksymalna odległość między dowolnymi dwoma ⊠⊠⊠⊠⊠⊠⊠⊠}.
dla każdej cząsteczki wykonaj
  $x_{1} = rand(x^{min}, x^{max})$ i $x_{2} = rand(x^{min}, x^{max})$
  $v_{1} = rand(v_1^{min}, v_1^{max})$ i $v_{2} = rand(v_2^{min}, v_2^{max})$
koniec dla
while iteration ≤ MaxIterations do
  for each particle do
    Ocena cząsteczki
    1. Oblicz $k$ dla punktu danych $ID = x_1$
    2. Oblicz funkcję fitness zgodnie z równaniem (4), używając $k$ i $r = x_2$
    Aktualizacja najlepszych wartości cząstek $x^{best}$, $x^{best}$
    Aktualizacja najlepszego $x$ roju $x_1^{gbest}$, $x_2^{gbest}$
  koniec dla
  dla każdej cząstki do
zakończone.

Oblicz prędkość i pozycję zgodnie z równaniem ( 3) i równaniem (2).
end for
end while
Używając $r = x^{gbest}$ oblicz$^{k}$ dla wszystkich punktów danych Posortuj punkty

1 $r$

## 3.4 Dyskusja

Inną ważną kwestią, z którą boryka się outPSO, jest gradient kondycji przestrzeni poszukiwań. Aby zmotywować cząstki do poruszania się w kierunku obiecujących regionów, w których najprawdopodobniej istnieją najlepsze rozwiązania, domena wyszukiwania powinna mieć gradient w krajobrazie kondycji. Krajobraz z płaskim poziomem kondycji i zawierający tylko skoki o wysokiej kondycji spowoduje, że PSO będzie nadal oscylować bez ustalania dobrego rozwiązania, a tym samym trudno będzie uzyskać zbieżność.

Przestrzeń poszukiwań nowego algorytmu outPSO jest określana przez indeks punktu, który jest wartością całkowitą, oraz *r*. W ten sposób outPSO jest powiązany ze sposobem indeksowania danych. W przypadku zestawów danych, które mają zostać przetestowane, outPSO nie napotka problemu związanego z kwestią gradientu. Ponadto nie jest wymagane znalezienie dokładnie optymalnego punktu, który ma minimalną wartość$^{k}$ . Punkty zbliżone do optymalnych będą wystarczające, ponieważ punkty są uporządkowane, a wartości odstające będą miały najniższy współczynnik$^{k}$ , nawet jeśli uporządkowanie wartości odstających można poprawić. Oczekujemy więc, że algorytm ten będzie działał dobrze w przypadku wykrywania wartości odstających.

$r$

$r$

# 4. EKSPERYMENTY

## 4.1 Wybór zestawu danych

Ocena metod wykrywania wartości odstających stanowi pewną trudność, ponieważ istnieją różne definicje wartości odstających, a różni eksperci dziedzinowi często mają różne opinie na temat tego, czy wykryty przypadek jest prawdziwą wartością odstającą, czy nie. W tym artykule używamy dwóch sposobów tworzenia testowych zestawów danych do eksperymentów. Pierwszym z nich jest użycie powszechnie akceptowanych "benchmarków". Drugim jest użycie istniejącego niezrównoważonego zbioru danych klasyfikacji, w którym niektóre lub wszystkie instancje z klasy mniejszości są wybierane jako odstające, podczas gdy instancje z klasy mniejszości są wybierane jako odstające.

ze wszystkich lub niektórych instancji z klasy (klas) większościowej jako grupy normalnej. Ponieważ istnieje tak wiele sprzecznych "punktów odniesienia" w tej dziedzinie, drugi sposób jest również szeroko stosowany w istniejących pracach [8]. Jednak w niektórych zbiorach danych niektóre elementy klas większościowych są tak różne, że można je wykryć jako wartości odstające.

Podążając za powyższymi dwoma sposobami, wybraliśmy pięć zestawów danych jako środowisko testowe: *zestaw danych Hocky*, *Wisconsin Breast Can- cer (Original)*, *Wisconsin Breast Cancer (Diagnostic)*, *zestaw danych Yeast* i zestaw danych *Shuttle* [14].

## 4.2 Konfiguracje eksperymentu

Algorytm outPSO jest oceniany przy wielkości populacji ustawionej na 30 cząstek i maksymalnej liczbie iteracji wynoszącej 1000. Cząstki są połączone w topologii pierścienia. Współczynnik zwężenia $\chi$ *jest ustawiony* na 0,729, $c1$ i $c2$ w równaniu 1 są ustawione na 2,02. Te wartości parametrów są ustawiane na podstawie wspólnych ustawień [16] i szybkich eksperymentów empirycznych. Każdy eksperyment jest powtarzany dla 100 niezależnych przebiegów i podawane są średnie wyniki.

Porównujemy outPSO z algorytmem LOF z *MinPtn* ustawiona na 40.

Miara odległości euklidesowej jest używana do obliczania odległości między dowolnymi dwoma punktami/instancjami danych. Wartości atrybutów we wszystkich zestawach danych są normalizowane/skalowane do [0, 1] zgodnie z równaniem 5.

$$f_i = \frac{(f_i \boxtimes f)^{min}_i}{(f^{max}_i \boxtimes f^{min}_i)} \quad (5)$$

gdzie $f_i$ jest wartością atrybutu dla konkretnej instancji, a $f^{max}_i$ i $f^{min}_i$ są wartościami maksymalną i minimalną tego atrybutu w całym zbiorze danych.

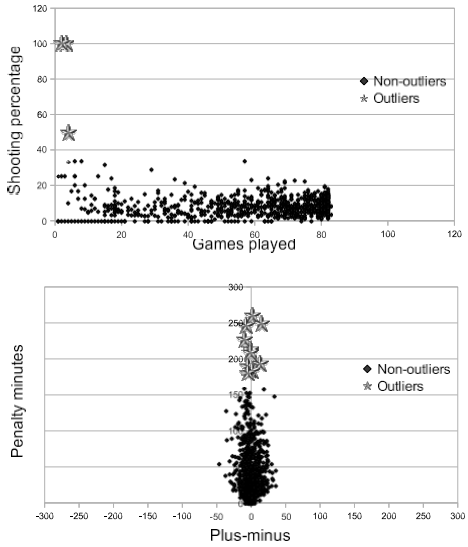
## 4.3 Wyniki: Zestaw danych hokejowych

Zbiór danych National Hockey League (NHL) został wykorzystany jako "punkt odniesienia" w kilku artykułach dotyczących wykrywania wartości odstających [11, 18]. Wykorzystano statystyki NHL 2003-2004 uzyskane ze strony internetowej NHL [13]. Te same statystyki zostały wykorzystane w [18]. Zbiór danych zawiera 916 wpisów.

Przeprowadzono dwa testy. Pierwszy test wyszukuje wartości odstające na podstawie trzech atrybutów: *rozegranych meczów, strzelonych bramek* i *procentu strzałów*. Wyniki przedstawiono w tabeli 1. Test ten jest stosunkowo łatwy, ponieważ gracze odstający różnią się od innych graczy (grupa normalna) pod względem wartości mediany trzech atrybutów. Nasza nowa metoda outPSO osiągnęła taką samą rangę zawodników odstających jak metoda LOF [11], a także metoda Yaling [18].

Drugi test polega na wykryciu wartości odstających na podstawie trzech różnych atrybutów: *zdobytych punktów, statystyk plus-minus* i *minut karnych*. Trzy najlepsze wartości odstające są identyczne dla algorytmów out- PSO i LOF, podczas gdy czwarta wartość odstająca wykryta przez outPSO zajęła trzecie miejsce w rankingu Yalinga [18].

W związku z tym podejście PSO osiągnęło co najmniej tak dobre wyniki, jak metody LOF i Yaling w dwóch testach. Ponadto PSO ma możliwość automatycznego wybierania ważnych atrybutów ze zbiorów danych. Rys. 4 wizualizuje dwa zestawy danych z odstającymi "ważnymi" (nie wszystkimi) atrybutami. W sekcji 3.4 wspomniano, że dla PSO w celu zlokalizowania regionów o dobrej kondycji kluczowy jest gradient kondycji. Dwa poprzednie testy są przeprowadzane na zestawach danych bez wcześniejszego porządkowania. Powtarzamy testy 100 razy z losowo zmienianymi pozycjami wartości odstających za każdym razem, aby




Rysunek 4: Wartości odstające wykryte w zbiorze danych NHL.

aby ponownie je wykryć. W przypadku obu testów outPSO jest w stanie znaleźć z prawdopodobieństwem 95% dowolny z trzech najlepszych punktów odstających jako najlepszy punkt. Gdy outPSO jest uruchamiane dwukrotnie, po pierwszym uruchomieniu punkty danych są uporządkowane, outPSO jest w stanie znaleźć dowolny punkt odstający.

najlepszych wartości odstających z prawdopodobieństwem 100%. Dzieje się tak dlatego, że uporządkowanie danych zapewnia lepszy gradient do wykonania analizy.

wyszukiwanie. W rzeczywistości, nawet bez zamawiania, nowa aplikacja PSO-

podejście nadal działało dobrze jako schemat kodowania cząstek

wziął pod uwagę krajobraz fitness.

zmienić krajobraz fitness i zbadać, czy outPSO jest w stanie

## 4.4 Wyniki: Wisconsin Breast Cancer Dataset (oryginalny) (WBCDO)

Zbiór danych WBCDO zawiera 699 instancji z dziewięcioma atrybutami [14]. Każdy rekord jest oznaczony jako łagodny (458 lub 65,5%) lub złośliwy (241 lub 34,5%). Zbiór danych wykrywania wartości odstających jest tworzony przez wybranie wszystkich 458 łagodnych rekordów (jako grupy normalnej) i pierwszych 10 złośliwych rekordów (jako wartości odstających). Ocena opiera się na liczbie złośliwych rekordów zajmujących 10 najlepszych pozycji.

Nowy algorytm outPSO jest w stanie wylistować 6 złośliwych rekordów, podczas gdy LOF nie wylistował żadnego złośliwego rekordu w pierwszej dziesiątce. Aby potwierdzić te wyniki, eksperyment powtórzono 100 razy z losowo wybranymi złośliwymi przykładami. Średnia liczba wartości odstających wykrytych przy użyciu outPSO wyniosła (5,85±1,17). Jednak algorytm LOF nie był w stanie wymienić żadnego z przykładów złośliwych na 10 najwyższych pozycjach.

Jednak gdy *MinPtn* wzrosło z 40 do 80, wydajność LOF poprawiła się, uzyskując średnio (4,55 ± 0,69), chociaż nadal jest to statystycznie istotnie gorsze niż metoda outPSO za pomocą standardowego testu T / Z (na poziomie ufności 95%). Zwiększanie *MinPtn* jest jednak kosztowne obliczeniowo. W tym przypadku LOF poświęcił średnio 15 sekund na zgłoszenie wartości odstających, podczas gdy nowa metoda outPSO wymagała mniej niż 0,5 sekundy czasu procesora, czyli 30 razy krócej niż metoda LOF.

## 4.5 Wyniki: Wisconsin Breast Cancer Dataset (Diagnostic) (WBCDD)

Ten zestaw danych zawiera 569 rekordów z 30 atrybutami [14]. Liczba wystąpień dla klasy *łagodnej* wynosi 357, a dla klasy

Tabela 1: Wykrywanie wartości odstających na zestawie danych Hockey, Test 1

| Ranga PSO | LOF Rank [4] | Yaling Rank[18] | Gracz | Rozegrane gry | Zdobyte bramki | Procent strzelców |
|---|---|---|---|---|---|---|
| 1 | 1 | 1 | Milan Michalek | 2 | 1 | 100 |
| 2 | 2 | 2 | Pat Kavanagh | 3 | 1 | 100 |
| 3 | 3 | 3 | Lubomir Sekeras | 4 | 1 | 50 |
| minimum | | | | 1 | 0 | 0 |
| mediana | | | | 57 | 6.4 | 6.6 |
| maksimum | | | | 83 | 41 | 100 |

Tabela 2: Wykrywanie wartości odstających w zbiorze danych Hockey, Test 2

| Ranga PSO | Ranga LOF | Yaling Rank[18] | Gracz | Zdobyte punkty | Plus-minus | Minuty karne |
|---|---|---|---|---|---|---|
| 1 | 1 | 1 | Sean Avery | 28 | 2 | 261 |
| 2 | 2 | 2 | Chris Simon | 28 | 15 | 250 |
| 3 | 3 | - | Krzysztof Oliwa | 5 | -8 | 247 |
| 4 | 6 | 3 | Jody Shelley | 6 | -10 | 228 |
| 5 | 8 | - | Donald Brashear | 13 | -1 | 212 |
| minimum | | | | 0 | -46 | 0 |
| mediana | | | | 12 | -1 | 26 |
| maksimum | | | | 94 | 35 | 261 |

dla klasy *złośliwej* wynosi 212. Zbiór danych do wykrywania wartości odstających w tym przypadku składa się z 357 instancji z klasy *łagodnej* (jako normalnej) i pierwszych 10 instancji z klasy *złośliwej* jako wartości odstających. Ze wszystkich 100 niezależnych przebiegów eksperymentu algorytm LOF poprawnie zgłosił
Średnio 5,21 ± 1,04 wartości odstających w 10 najlepszych pozycjach, podczas gdy outPSO odnotowało 5,23 ± %0,95 z 10 wartości odstających. W tym przypadku nowy algorytm PSO jest nieco lepszy niż metoda LOF, ale różnica nie jest statystycznie istotna dla standardowego testu Z na poziomie ufności 95%.

## 4.6 Wyniki: Zestaw danych drożdży

Ten zbiór danych składa się z 1484 instancji dla dziewięciu klas [14]. Każda instancja ma osiem atrybutów. Klasa ERL ma tylko pięć przykładów. Używamy tej klasy jako wartości odstających w stosunku do pierwszych trzech klas (CYT, NUC, MIT), które składają się z 1136 instancji (grupa normalna). Nowy algorytm outPSO z powodzeniem zidentyfikował wszystkie pięć wartości odstających dla klasy ERL na pierwszych pięciu pozycjach. Jednak tych pięciu członków nie zajmowało żadnej pozycji w pierwszej piątce listy LOF. Tabela 4.6 przedstawia 16 najlepszych wartości odstających wykrytych przez outPSO i odpowiadający im ranking LOF. Wśród 10 najlepszych pozycji metod LOF tylko jedna z pięciu wartości odstających (P5) została poprawnie wykryta, a wszystkie inne wartości odstające (P1 - P4) zostały pominięte, które zostały sklasyfikowane na pozycjach 13-16 (nieprawidłowo wykryte jako należące do grupy normalnej). Sugeruje to, że algorytm PSO przewyższa metodę LOF w tym zestawie danych.

Aby uzyskać wrażenie, dlaczego te punkty danych zostały prawidłowo wykryte jako wartości odstające przez LOF, rys. 5 pokazuje pozycje tych wartości odstających w odniesieniu do innych punktów (cecha 5 vs cecha 6, powyżej) i (cecha 1 vs cecha 3, poniżej). W tych dwóch podprzestrzeniach wiele wartości odstających (P6-P16) było w rzeczywistości dość daleko od normalnych punktów/instancji, a zatem łatwo je pomylić z wartościami odstającymi. Być może dlatego metoda LOF nieprawidłowo wykryła wiele z nich jako wartości odstające. Z drugiej strony nowa metoda PSO ma możliwość automatycznego wyboru ważnych atrybutów/cech, dzięki czemu ma większą zdolność do wykrywania prawidłowych wartości odstających z mylących przypadków.

Tabela 3: Wartości odstające wykryte w zestawie danych Yeast

| Punkty | Klasa | Ranga outPSO | Ranga LOF |
|---|---|---|---|
| P1 | ERL | 1 | 15 |
| P2 | ERL | 2 | 16 |
| P3 | ERL | 3 | 13 |
| P4 | ERL | 4 | 12 |
| P5 | ERL | 5 | 10 |
| P6 | NUC | 6 | 14 |
| P7 | NUC | 7 | 5 |
| P8 | MIT | 8 | 6 |
| P9 | NUC | 9 | 8 |
| P10 | CYT | 10 | 9 |
| P11 | CYT | 11 | 17 |
| P12 | CYT | 12 | 18 |
| P13 | MIT | 13 | 3 |
| P14 | CYT | 14 | 4 |
| P15 | CYT | 15 | 1 |
| P16 | CYT | 16 | 2 |

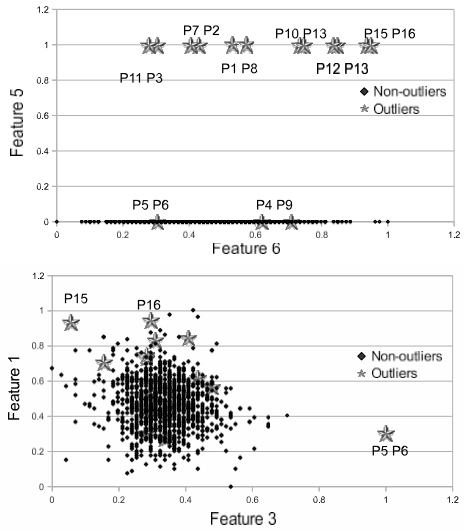
## 4.7 Zestaw danych Shuttle

Aby porównać wydajność obliczeniową nowego algorytmu out- PSO i algorytmu LOF, użyliśmy większego zbioru danych z większą liczbą przykładów, ponieważ oba algorytmy mogą być szybkie na małych zbiorach danych, co nie rozróżnia wyraźnie prędkości. Wykorzystano tutaj zbiór danych shuttle z dziewięcioma przykładami [14]. Do wykrywania wartości odstających wykorzystano zestaw testowy zawierający 14500 przykładów. Oba algorytmy wykryły wartość odstającą *ID* = 11750.

Jeśli chodzi o czas wykonania, algorytm outPSO potrzebował średnio 170 sekund w 100 niezależnych przebiegach eksperymentu, podczas gdy LOF potrzebował średnio 936 sekund. Wyraźnie widać, że nowy algorytm outPSO jest znacznie szybszy niż LOF w tym zestawie danych. Potwierdza to naszą wczesną hipotezę, że nowy algorytm PSO jest bardziej wydajny, ponieważ wyszukuje tylko obiecujące regiony, zamiast przeszukiwać wszystkich sąsiadów.

# 5. WNIOSKI

Celem niniejszego artykułu było opracowanie nowego podejścia opartego na PSO do wykrywania wartości odstających przy użyciu wspólnej odległości.




Rysunek 5: Przykład fałszywie dodatnich wartości odstających w zbiorze danych Yeast.

oparte na środkach. XS Cel ten został pomyślnie osiągnięty. Wyniki pokazują, że nowe podejście oparte na PSO osiągnęło znacznie lepszą wydajność niż metoda LOF w większości tych zestawów danych i porównywalną lub nieco lepszą wydajność w niektórych zestawach danych. Ponadto nowa metoda oparta na PSO jest bardziej wydajna niż powszechnie stosowana metoda LOF.

W porównaniu z niektórymi metodami opartymi na odległości [11, 4], nowa metoda PSO nie wymaga od użytkowników ręcznego określania kluczowych parametrów miary odległości związanych ze zbiorami danych, chociaż parametry ewolucyjne nadal muszą być ustawione.

Struktura oparta na PSO opracowana w tej pracy różni się od istniejących rozwiązań ewolucyjnych do wykrywania wartości odstających, w których PSO lub GA były używane głównie do selekcji cech, a wybrane cechy są wykorzystywane do wykrywania wartości odstających przy użyciu innych (nieewolucyjnych) metod. To nowe podejście integruje zdolność selekcji cech w całej strukturze, która może *bezpośrednio* i *automatycznie* wykrywać wartości odstające z określonego zbioru danych.

## 6. ODNIESIENIA


[1] C. C. Aggarwal i P. S. Yu. Wykrywanie wartości odstających dla danych wielowymiarowych. In *Proceedings of the 2001 ACM SIGMOD International Conference on Management of Data*, strony 37-46, 2001.

[2] M. Agyemang i C. Ezeife. Solidny schemat wykrywania wartości odstających dla dużych zbiorów danych. In *6th Pacific-Asia Conference on Knowledge Discovery and Data Mining*, strony 6-8, 2001.

[3] I. Ben-Gal. *Maimon O. i Rockach L.:Data Mining and Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers*. Kluwer Academic Publishers, 2005.

[4] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander. Lof: identyfikacja lokalnych wartości odstających na podstawie gęstości. *SIGMOD Rec.*, 29(2):93-104, 2000.

[5] V. Chandola, A. Banerjee i V. Kumar. Wykrywanie anomalii: przegląd. *ACM Comput. Surv.*, 41(15):1-58, lipiec 2009.

[6] K. D. Crawford i R. L. Wainwright. Zastosowanie algorytmy genetyczne do wykrywania wartości odstających. In *Proceedings of the 6th International Conference on Genetic Algorithms*, strony 546-550, 1995.

[7] D. Hawkins. *Identification of Outliers*. Chapman and Hall, 1980.

[8] S. Hawkins, H. He, G. Williams i R. Baxter. Wykrywanie wartości odstających przy użyciu sieci neuronowych replikatorów. In *Proc. of the Fifth Int. Conf. and Data Warehousing and Knowledge Discovery (DaWaK02*, strony 170-180, 2002.

[9] V. Hodge i J. Austin. Przegląd metodologii wykrywania wartości odstających. *Artif. Intell. Rev.*, 22(2):85-126, październik 2004.

[10] J. Kennedy i R. C. Eberhart. Optymalizacja rojem cząstek. In *IEEE Int. Conf. on Neural Networks*, strony 1942-1948, 1995.

[11] E. M. Knorr, R. T. Ng i V. Tucakov. Wartości odstające oparte na odległości: Algorytmy i zastosowania. *VLDB Journal: Very Large Data Bases*, 8(3-4):237-253, 2000.

[12] H.-P. Kriegel, M. S hubert i A. Zimek. Wykrywanie wartości odstających na podstawie kąta w danych wielowymiarowych. In *Proceeding of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, strony 444-452, 2008.

[13] T. N. H. League. http://www.nhl.com/.

[14] UCI Repository of Machine Learning Databases. http://archive.ics.uci.edu/ml/datasets.html

[15] Y. Li i H. Kitagawa. Db-outlier detection by example in high dimensional datasets. In *SWOD '07: Proceedings of the 2007 IEEE International Workshop on Databases for Next Generation Researchers*, strony 73-78, 2007.

[16] C. M. Rój i królowa: w kierunku deterministycznej i adaptacyjnej optymalizacji roju cząstek. In *IEEE Congress on Evolutionary Computation*, tom 2, strony 1951-1957, 1999.

[17] S. Papadimitriou, H. Kitagawa, P. B. Gibbons, i C. Faloutsos. Loci: szybkie wykrywanie wartości odstających przy użyciu lokalnej całki korelacji. In *Proceedings of 19th International Conference on Data Engineering*, strony 315-326, maj 2003.

[18] Y. Pei, O. Zaiane i Y. Gao. Efektywny oparte na referencjach podejście do wykrywania wartości odstających w dużych zbiorach danych. In *Sixth International Conference on Data Mining*, strony 478-487, grudzień 2006.

[19] S. Ramaswamy, R. Rastogi i K. Shim. Efficient algorithms for mining outliers from large data sets. *SIGMOD Rec.*, 29(2):427-438, 2000.

[20] J. Tang, Z. Chen, A. W.-C. Fu i D. Cheung. Lsc-mine: Algorytm do wydobywania lokalnych wartości odstających. *In 15 Międzynarodowa Konferencja Stowarzyszenia Zarządzania Zasobami Informacyjnymi (IRMA)*, maj 2004.

[21] D. Ye1 i Z. Chen. Nowy algorytm wysokowymiarowego wykrywania wartości odstających oparty na constrained particle swarm intelligence. *Lecture Notes in Computer Science*, 5009/2008:516-523, maj 2008.

[22] K. Zhang, M. Hutter i H. Jin. Tytuł: Nowe lokalne podejście do wykrywania wartości odstających oparte na odległości dla rozproszonych danych ze świata rzeczywistego. In *Proc. 13th Pacific-Asia Conf. on Know. Discov. and Data Mining*, s. 813-822, 2009.