Politechnika Śląska
Wydział Automatyki, Elektroniki i Informatyki
Instytut Informatyki

Autoreferat i streszczenie rozprawy doktorskiej

Genetic and memetic algorithms for selection of training sets for support vector machines

Jakub Nalepa

Promotor: dr hab. inż. Michał Kawulok

Gliwice, 2016

Spis treści

Rozdział 1

Streszczenie rozprawy doktorskiej

1.1 Wprowadzenie

Maszyna wektorów podpierających (ang. *support vector machine*, SVM) jest klasyfikatorem nadzorowanym [3], który znalazł zastosowanie w rozwiązywaniu wielu zadań związanych z rozpoznawaniem wzorców [2, 4, 5, 7, 8, 15–17]. Trening klasyfikatora polega na wyznaczeniu hiperpłaszczyzny decyzyjnej separującej dane (tj. wektory cech) należące do dwóch klas w zbiorze treningowym $\boldsymbol{T}$. W przedstawionej pracy rozważany jest problem klasyfikacji dwuklasowej. Położenie hiperpłaszczyzny separującej jest zdefiniowane przez pewien reprezentatywny podzbiór zbioru $\boldsymbol{T}$, nazywany *wektorami podpierającymi* (ang. *support vectors*)*. Hiperpłaszczyzna jest następnie wykorzystana do klasyfikacji nowych danych, które nie zostały użyte w czasie treningu.

Problem treningu klasyfikatora SVM jest zagadnieniem programowania kwadratowego z ograniczeniami, o złożoności czasowej $O(t^3)$ i pamięciowej $O(t^2)$, gdzie t oznacza liczbę wektorów w zbiorze $\boldsymbol{T}$. Dla bardzo dużych zbiorów danych (powszechnych w wielu dziedzinach nauki i przemysłu, np. w bioinformatyce [13]), złożony proces treningu staje się istotną wadą klasyfikatora i utrudnia jego wykorzystanie. Dodatkowo liczba wyznaczonych wektorów podpierających (oznaczana jako s) wpływa na czas późniejszej klasyfikacji, której złożoność czasowa wynosi $O(s)$. Liczba wektorów podpierających jest w praktyce proporcjonalna do wielkości zbioru treningowego. Oznacza to, że duża wielkość zbioru $\boldsymbol{T}$ nie tylko wydłuża czas treningu klasyfikatora, ale także pośrednio wpływa na czas klasyfikacji. Zmniejszenie liczby wyznaczonych wektorów podpierających (bez jednoczesnego wyraźnego zmniejszenia jakości klasyfikacji) pozwala więc na skrócenie czasu klasyfikacji oraz na zastosowanie klasyfikatora SVM w aplikacjach dzia-

*Oznacza to, że możliwe jest takie wyłonienie podzbioru zbioru treningowego, dla którego otrzymane wektory podpierające są takie same jak dla pełnego zbioru treningowego.

łających w czasie rzeczywistym, nawet w przypadku dużych zbiorów danych. Warto wspomnieć, że w zbiorze treningowym mogą pojawić się wektory posiadające błędną etykietę. Wektory te powinny zostać usunięte ze zbioru $\boldsymbol{T}$ przed przeprowadzeniem treningu klasyfikatora SVM.

Zagadnienie treningu klasyfikatora SVM w przypadku bardzo dużych zbiorów danych jest znane w świecie naukowym. W literaturze można wyróżnić dwa główne rodzaje metod rozwiązywania tego problemu. Pierwsza grupa zawiera metody, których celem jest zoptymalizowanie i przyspieszenie samego procesu treningu klasyfikatora. Do drugiej grupy należą algorytmy, których zadaniem jest taki dobór podzbioru pełnego zbioru treningowego $\boldsymbol{T}$ (podzbiór ten oznaczony jest jako $\boldsymbol{T'}$), aby w tym podzbiorze znalazły się wektory potencjalnie „istotne", tj. takie, które z dużym prawdopodobieństwem zostaną wybrane jako wektory podpierające w czasie treningu. Dobór odpowiedniego podzbioru $\boldsymbol{T'}$ (zawierającego wyraźnie mniej wektorów t' niż pełny zbiór $\boldsymbol{T}$, tj. $t' \ll t$) pozwala nie tylko na przyspieszenie treningu klasyfikatora SVM, ale także na zmniejszenie liczby wyznaczonych wektorów podpierających oraz wyeliminowanie tych wektorów ze zbioru $\boldsymbol{T}$, które negatywnie wpływają na położenie hiperpłaszczyzny separującej [11, 12]. Nie jest to możliwe w przypadku metod należących do pierwszej grupy, w których wykorzystywany jest pełny zbiór $\boldsymbol{T}$ podczas treningu klasyfikatora SVM.

Algorytmy doboru zredukowanego zbioru treningowego $\boldsymbol{T'}$ można najogólniej podzielić na trzy grupy: (i) metody oparte o analizę geometrii danych w zbiorze $\boldsymbol{T}$, (ii) metody statystyczne, oraz (iii) inne, do których zaliczają się m.in. algorytmy aktywnego uczenia oraz próbkowania losowego. Głównymi celami tworzenia nowych i ulepszonych algorytmów do doboru zredukowanego zbioru treningowego $\boldsymbol{T'}$ są:

- Umożliwienie przeprowadzenia procesu treningu klasyfikatora SVM nawet dla bardzo dużych zbiorów danych.

- Polepszenie jakości klasyfikacji poprzez usunięcie tych wektorów z pełnego zbioru $\boldsymbol{T}$, które mogą negatywnie wpłynąć na jakość wyznaczonej hiperpłaszczyzny.

- Zmniejszenie liczby wyznaczonych wektorów podpierających, w celu przyspieszenia klasyfikacji.

Warto zauważyć, że algorytmy ewolucyjne nie były dotąd intensywnie stosowane do wyboru zbioru $\boldsymbol{T'}$ (pomimo tego, że znalazły zastosowanie w rozwiązywaniu wielu innych złożonych problemów obliczeniowych [10]). Początkowe obserwacje wskazujące na to, że algorytmy genetyczne mogą być z powodzeniem zastosowane do tego celu, zostały opisane w publikacji [6].

1.2 Tezy i cele pracy

Dwie tezy rozprawy zostały sformułowane w języku angielskim:

1. **Applying adaptation techniques in genetic algorithms for selection of training sets for support vector machines allows for:**

   (a) **delivering better refined training sets, and**

   (b) **improving the process of retrieving refined training sets,**

   compared with the genetic algorithms which do not involve the adaptation.

2. **Applying memetic algorithms, which are aimed at exploiting knowledge (extracted before the evolutionary optimization, attained during the search, or both) concerned with the vectors in the training set, for selection of training sets for support vector machines allows for retrieving better refined training sets compared with those elaborated using genetic algorithms, and other state-of-the-art techniques.**

Tłumaczenie powyższych tez na język polski jest następujące:

1. **Zastosowanie technik adaptacji w algorytmach genetycznych dla doboru zbiorów treningowych dla maszyny wektorów podpierających pozwala na:**

   (a) **uzyskanie lepszych zredukowanych zbiorów treningowych oraz na**

   (b) **polepszenie procesu doboru zbiorów treningowych,**

   w porównaniu z innymi algorytmami genetycznymi, w których nie wykorzystano adaptacji.

2. **Zastosowanie algorytmów memetycznych, których celem jest wykorzystanie dostępnej wiedzy (wyekstrahowanej przed optymalizacją ewolucyjną lub/i uzyskaną podczas ewolucji) dotyczącej wektorów ze zbioru treningowego, do doboru zbiorów treningowych dla maszyny wektorów podpierających, pozwala na uzyskanie lepszych zredukowanych zbiorów treningowych w porównaniu z tymi, które zostały otrzymane przy użyciu algorytmów genetycznych i innych algorytmów znanych z literatury.**

Warto zauważyć, że ocena jakości zredukowanych zbiorów treningowych nie jest trywialna. Jak wspomniano wcześniej, „pożądany" zredukowany zbiór treningowy powinien pozwalać na polepszenie jakości działania klasyfikatora, zmniejszenie liczby wyznaczonych wektorów podpierających oraz na umożliwienie przeprowadzenia treningu (w przypadku bardzo dużych zbiorów danych).

W celu zweryfikowania powyższych tez, zdefiniowano następujące **główne cele** rozprawy:

1. Zaprojektowanie i zaimplementowanie algorytmu genetycznego dla doboru zbioru treningowego dla klasyfikatora SVM.

2. Zaprojektowanie i zaimplementowanie adaptacyjnych algorytmów genetycznych dla doboru zbioru treningowego dla klasyfikatora SVM, wykorzystujących zróżnicowane techniki adaptacji.

3. Zaprojektowanie i zaimplementowanie algorytmu memetycznego dla doboru zbioru treningowego dla klasyfikatora SVM, wykorzystującego wiedzę o wektorach należących do zbioru treningowego zdobytą podczas ewolucji.

4. Zaprojektowanie i zaimplementowanie adaptacyjnego algorytmu memetycznego wykorzystującego dodatkową analizę geometrii danych dla doboru zbioru treningowego dla klasyfikatora SVM, tj. algorytmu wykorzystującego wiedzę o wektorach należących do zbioru treningowego zdobytą podczas ewolucji oraz uzyskaną podczas przetwarzania wstępnego (preprocessingu) zbioru treningowego (przed ewolucją).

5. Walidacja eksperymentalna zaproponowanych algorytmów przy użyciu zbiorów danych wzorcowych (benchmarkowych), rzeczywistych i sztucznie wygenerowanych.

6. Zweryfikowanie wpływu zaproponowanych rozwiązań algorytmicznych na jakość wyników otrzymywanych za pomocą algorytmów, w których rozwiązania te zostały wykorzystane.

7. Porównanie wyników (tj. jakości klasyfikacji oraz liczby wektorów podpierających) otrzymywanych za pomocą zaproponowanych algorytmów i innych algorytmów znanych z literatury.

Powyższe główne cele rozprawy zostały uzupełnione poniższymi **celami drugorzędnymi**:

1. Zwizualizowanie zredukowanych zbiorów treningowych (wraz z wyznaczonymi wektorami podpierającymi) otrzymanych przy użyciu zaproponowanych algorytmów oraz tych znanych z literatury.

2. Zweryfikowanie statystycznej istotności otrzymywanych wyników.

1.3 Zaproponowane algorytmy

W przedstawionej rozprawie opisano **pięć** zaproponowanych algorytmów ewolucyjnych dla doboru zbioru treningowego dla klasyfikatora SVM:

- Algorytm genetyczny (ang. *genetic algorithm*, GASVM),
- Adaptacyjny algorytm genetyczny (ang. *adaptive genetic algorithm*, AGA),
- Dynamicznie adaptacyjny algorytm genetyczny (ang. *dynamically adaptive genetic algorithm*, DAGA),
- Algorytm memetyczny (ang. *memetic algorithm*, MASVM), oraz
- Adaptacyjny algorytm memetyczny wykorzystujący dodatkową analizę geometrii danych w zbiorze $\boldsymbol{T}$ (PCA^{2}MA).

Zaprojektowanie oraz zaimplementowanie powyższych algorytmów pozwoliło na zrealizowanie czterech (z siedmiu) głównych celów rozprawy (cele 1–4).

1.3.1 Algorytmy genetyczne

W algorytmie genetycznym (GASVM), populacja rozwiązań[†] o wielkości N, w której każdy osobnik p_i (chromosom) reprezentuje pewien podzbiór zbioru $\boldsymbol{T}$ zawierający t' wektorów, ewoluuje w czasie. Podczas ewolucji, osobniki są wybierane do krzyżowania w procesie selekcji, następnie są krzyżowane, a osobnik potomny jest dodatkowo poddawny operacji mutacji oraz kompensacji, mającej na celu dołączenie nowych wektorów ze zbioru $\boldsymbol{T}$ do osobnika potomnego w przypadku, gdy liczba wektorów w tym osobniku jest mniejsza niż t'. Wielkość osobników, tj. liczba wektorów w każdym zredukowanym zbiorze treningowym, jest stała w całej populacji i musi zostać określona przed uruchomieniem algorytmu genetycznego. Jakość osobników określana jest na podstawie wartości funkcji przystosowania (η), która kwantyfikuje jakość klasyfikacji klasyfikatora SVM, którego trening został przeprowadzony przy użyciu zredukowanego zbioru treningowego, reprezentowanego przez danego osobnika. Funkcja ta może zostać określona jako pole powierzchni pod krzywą ROC (ang. *receiver operating characteristic*) wyznaczoną dla zbioru treningowego $\boldsymbol{T}$, lub procent poprawnie sklasyfikowanych wektorów ze zbioru $\boldsymbol{T}$. W algorytmie GASVM wykorzystano różne schematy selekcji rozwiązań do krzyżowania.

Zaprojektowanie i zaimplementowanie algorytmu GASVM pozwoliło na **zrealizowanie pierwszego z siedmiu głównych celów przedstawionej rozprawy**.

[†]Początkowa populacja rozwiązań generowana jest losowo.

Istotnym problemem dotyczącym algorytmu GASVM była trudność doboru odpowiedniej wielkości zredukowanych zbiorów treningowych (tj. wielkości osobnika w populacji). Kosztowny obliczeniowo proces strojenia tej wartości musiał być przeprowadzony dla każdego nowego zbioru treningowego, a niepoprawnie dobrana wartość parametru t' wyraźnie wpływała na jakość otrzymywanych wyników i czas optymalizacji. W adaptacyjnym algorytmie genetyczym (AGA), wielkość osobników w populacji jest dynamicznie uaktualniana w czasie ewolucji. Pozwala to na lepsze zbalansowanie eksploatacji i eksploracji przestrzeni rozwiązań, oraz na pominięcie procesu strojenia wartości t', który jest szczególnie uciążliwy w przypadku dużych zbiorów danych. Istotnym elementem algorytmu AGA jest nowa metoda oceny zróżnicowania populacji. W celu weryfikacji zróżnicowania populacji, przeprowadzana jest redukcja wymiarowości za pomocą analizy składowych głównych (ang. *principal component analysis*, PCA), a następnie wyznaczane są przedziały (osobno dla każdej klasy) w każdym z wymiarów po redukcji. Wektory obu klas są grupowane w ten sposób, żeby w każdym z przedziałów znajdowała się taka sama liczba wektorów ze zbioru $\boldsymbol{T}$. Każdy osobnik jest charakteryzowany przez histogram obrazujący przynależność poszczególnych wektorów do wyznaczonych przedziałów. Podobieństwo dwóch osobników w populacji jest następnie określane na podstawie podobieństwa dwóch charakteryzujących je histogramów. W przypadku małego zróżnicowania populacji, przeprowadzana jest operacja regeneracji, w której najlepsze osobniki z obecnej generacji są kopiowane, a pozostałe chromosomy są tworzone losowo. W algorytmie AGA wykorzystana została także adaptacyjna metoda krzyżowania osobników, oraz adaptacyjny schemat aktualizacji wielkości populacji podczas ewolucji.

W dynamicznie adaptacyjnym algorytmie genetycznym (DAGA), współczynnik zwiększania osobnika (tj. liczby wektorów w zredukowanym zbiorze treningowym) jest uaktualniany dynamicznie podczas ewolucji. Proces aktualizacji tego współczynnika pozwala na lepsze dostosowanie dynamiki wzrostu chromosomów (dla przykładu, szybszy wzrost zredukowanych zbiorów treningowych jest często bardzo korzystny dla dużych zbiorów danych, dla których liczba wektorów podpierających powinna być wyraźnie większa), oraz na szybsze wyznaczenie pożądanej wielkości zredukowanych zbiorów $\boldsymbol{T'}$. W algorytmie DAGA zdefiniowany został *współczynnik wypełnienia*, określony jako stosunek liczby wektorów podpierających s i liczby wszystkich wektorów w zredukowanym zbiorze treningowym. Jeśli wartość tego współczynnika jest wysoka, to uzasadniony jest szybszy wzrost wielkości zredukowanego zbioru (eksploracja). Jeżeli jednak niewielki odsetek wszystkich wektorów z danego zbioru wybieranych jest jako wektory podpierające, to wzrost jest spowolniony (w celu lepszej eksploatacji obecnej wartości t').

Zaprojektowanie i zaimplementowanie algorytmów AGA i DAGA pozwoliło na **zrealizowanie drugiego z siedmiu głównych celów przedstawionej rozprawy**.

1.3.2 Algorytmy memetyczne

Algorytmy memetyczne są technikami hybrydowymi, łączącymi ze sobą algorytmy ewolucyjne (stosowane do eksploracji przestrzeni rozwiązań) z algorytmami lokalnych ulepszeń (stosowanymi do eksploatacji pewnej części przestrzeni rozwiązań), które mogą wykorzystywać informacje uzyskane podczas ewolucji lub wyekstrahowane przed optymalizacją. W zaproponowanym algorytmie memetycznym do doboru zbiorów treningowych dla klasyfikatora SVM (MASVM), wykorzystane zostały informacje dotyczące wektorów wyznaczonych jako podpierające w czasie optymalizacji. Wektory te tworzą pulę potencjalnie istotnych i cennych wektorów ($\boldsymbol{P}$), które mogą zostać użyte do polepszenia innych chromosomów w populacji (podczas operacji *edukacji*), oraz do stworzenia osobników (tzw. *superosobników*, ang. *super individuals*) zawierających wyłącznie „cenne" wektory ze zbioru $\boldsymbol{P}$. Zastosowanie powyższych technik wraz z nowym, dwustanowym schematem selekcji rozwiązań do krzyżowania, którego celem jest zapewnienie odpowiedniej równowagi pomiędzy eksploracją i eksploatacją przestrzeni rozwiązań, pozwoliło na wyraźne zmniejszenie wielkości otrzymywanych zredukowanych zbiorów treningowych (przy jednoczesnym zapewnieniu wysokiej jakości klasyfikacji) oraz na skrócenie czasu optymalizacji.

Zaprojektowanie i zaimplementowanie algorytmu MASVM pozwoliło na **zrealizowanie trzeciego z siedmiu głównych celów przedstawionej rozprawy**.

We wszystkich wymienionych algorytmach ewolucyjnych, początkowa populacja rozwiązań była tworzona przy użyciu próbkowania losowego pełnego zbioru $\boldsymbol{T}$. W adaptacyjnym algorytmie memetycznym (PCA^{2}MA) zastosowano dodatkową analizę wstępną zbioru $\boldsymbol{T}$, mającą na celu wyznaczenie użytecznych wektorów przed ewolucją (tj. uzyskanie dodatkowych informacji dotyczących wektorów w zbiorze treningowym w procesie przetwarzania wstępnego – preprocessingu). Podczas przetwarzania wstępnego, tworzona jest zredukowana przestrzeń cech przy użyciu PCA, w której wyznaczane są przedziały w taki sposób, żeby zbiór treningowy został podzielony na równoliczne podzbiory (osobno dla każdej z klas). Następnie, dla każdego podzbioru wyznaczany jest wektor średni oraz analizowana jest odległość Mahalanobisa każdego wektora z przedziału do wektora średniego. Wektory w każdym przedziale są sortowane według tej odległości, a wektory o największych odległościach w każdym przedziale umieszczane są w puli wektorów-kandydatów (ang. *candidate vectors*), wykorzystywanych w różnych etapach algorytmu PCA^{2}MA (np. podczas tworzenia początkowej populacji rozwiązań oraz podczas procesu kompensacji). W algorytmie PCA^{2}MA zaproponowano bezparametryczny schemat adaptacji, który nie wymaga wyznaczenia wartości żadnych parametrów przed optymalizacją memetyczną. Podobnie jak w przypadku algorytmu MASVM, w algorytmie PCA^{2}MA wykorzystano pulę wektorów $\boldsymbol{P}$.

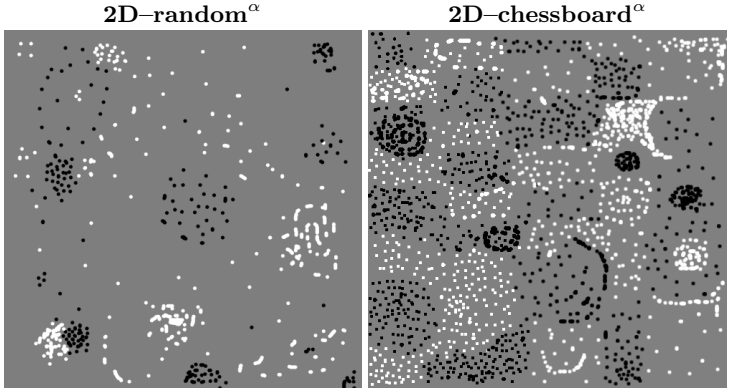
Zaprojektowanie i zaimplementowanie algorytmu PCA^{2}MA pozwoliło na **zrealizowanie czwartego z siedmiu głównych celów przedstawionej rozprawy**.

Zaproponowane algorytmy ewolucyjne zostały zbadane teoretycznie: przeanalizowano ich złożoność czasową oraz pamięciową. Przeprowadzono także teoretyczną analizę ich zbieżności.

1.4 Wyniki eksperymentalne

Wszystkie zaproponowane algorytmy ewolucyjne oraz algorytmy znane z literatury do doboru zredukowanych zbiorów treningowych dla klasyfikatora SVM zostały zaimplementowane w języku C++ i uruchomione na komputerze klasy PC (z procesorem Intel Xeon 3,2 GHz, oraz 16 GB pamięci RAM). Stworzenie odpowiedniego środowiska badawczego wymagało zaimplementowania konwertera danych dostępnych w repozytorium UCI, stworzenia programu komputerowego do automatycznego doboru wartości parametrów funkcji jądrowych dla klasyfikatora SVM oraz przetwarzania wsadowego, a także programu do wizualizacji otrzymywanych zredukowanych zbiorów treningowych i wektorów podpierających, oraz do analizy statystycznej otrzymywanych wyników.

Walidacja eksperymentalna zaimplementowanych metod doboru zbiorów treningowych była realizowana przy użyciu czterech zbiorów sztucznie wygenerowanych, przedstawiających punkty na płaszczyźnie i łatwych do wizualizacji (dwa przykładowe zbiory są przedstawione na Rysunku 1.1), dwóch zbiorów benchmarkowych z repozytorium UCI, oraz zbioru zawierającego dane rzeczywiste wyekstrahowane z bazy obrazów barwnych przedstawiających obszary ludzkiej skóry (zbiór ten jest zbiorem niezbalansowanym)[‡].



Rysunek 1.1: Dwa przykładowe sztucznie wygenerowane zbiory danych 2D – białe i czarne piksele oznaczają wektory należące do dwóch klas.

[‡]Repozytorium UCI jest dostępne pod adresem: `http://archive.ics.uci.edu/ml/`, natomiast zbiory sztucznie wygenerowane oraz zbiór zawierający dane rzeczywiste są dostępne pod adresem: `http://sun.aei.polsl.pl/~jnalepa/SVM/`.

Dobór zredukowanego zbioru treningowego dla klasyfikatora SVM może być traktowany jako wielokryterialny problem optymalizacyjny. Jego głównym celem jest znalezienie takiego zbioru $\boldsymbol{T'}$, który pozwoli na wyznaczenie hiperpłaszczyzny decyzyjnej, dla której jakość klasyfikacji (wyznaczona jako odsetek poprawnie sklasyfikowanych danych ze zbioru walidacyjnego lub jako pole pod krzywą ROC) jest jak najwyższa. Drugim kryterium jest zminimalizowanie liczby wyznaczonych wektorów podpierających w celu zmniejszenia czasu klasyfikacji. W przedstawionej rozprawie zdefiniowano funkcję jakości Q, pozwalającą na skwantyfikowanie jakości zredukowanych zbiorów treningowych otrzymanych przy użyciu analizowanych metod, na podstawie powyższych kryteriów:

$$Q(\eta_{\boldsymbol{V}}, s) = q \cdot \frac{\eta_{\boldsymbol{V}}}{\eta_{\boldsymbol{V}}^{\mathrm{B}}} + (1 - q) \cdot \frac{s^{\mathrm{B}}}{s}, \tag{1.1}$$

gdzie $\eta_{\boldsymbol{V}}$ jest polem pod krzywą ROC uzyskaną dla zbioru walidacyjnego, $\eta_{\boldsymbol{V}}^{\mathrm{B}}$ oznacza największą wartość pola (spośród wszystkich metod) pod krzywą ROC uzyskaną dla zbioru walidacyjnego, s^{B} oznacza najmniejszą liczbę wektorów podpierających (uzyskaną dla zredukowanych zbiorów treningowych otrzymanych przy użyciu analizowanych algorytmów), a q jest współczynnikiem ważności pierwszego kryterium ($0 < q \leq 1$). Statystyczna istotność otrzymywanych wyników została zweryfikowana przy użyciu testów Wilcoxona.

W celu rzetelnego i przekrojowego porównania analizowanych metod, wszystkie badania eksperymentalne zostały podzielone na następujące grupy: (i) analizę czułościową zaproponowanych algorytmów ewolucyjnych, (ii) porównanie zaproponowanych algorytmów ewolucyjnych, oraz (iii) porównanie metod ewolucyjnych z innymi algorytmami znanymi z literatury. Wizualizacje zbiorów treningowych oraz wektorów podpierających dla zbiorów sztucznie wygenerowanych zostały użyte do porównania jakościowego wyników otrzymywanych za pomocą wszystkich przeanalizowanych algorytmów do doboru zbioru treningowego (zbadano również, które wektory ze zbioru $\boldsymbol{T}$ są oznaczane jako podpierające, jeżeli pełny zbiór $\boldsymbol{T}$ jest użyty do treningu klasyfikatora SVM). W poniższych sekcjach podsumowano wyniki wymienionych badań eksperymentalnych.

1.4.1 Analiza czułościowa

W celu zweryfikowania wpływu zaproponowanych rozwiązań algorytmicznych na jakość otrzymywanych wyników, przeprowadzona została analiza czułościowa każdego z algorytmów ewolucyjnych. W algorytmie genetycznym (GASVM) zbadano wpływ zastosowanego schematu selekcji oraz wielkości osobników w populacji, w adaptacyjnym algorytmie genetycznym (AGA) przeanalizowano wpływ schematu adaptacji wielkości osobników podczas ewolucji oraz nowej metody krzyżowania rozwiązań. Warto zauważyć, że wyniki badań eksperymentalnych wyka-

zały, że w przypadku niektórych zbiorów ciągłe zwiększanie wielkości zredukowanych zbiorów $\boldsymbol{T'}$ powoduje pogorszenie jakości klasyfikacji (tj. istnieją wektory w zbiorze $\boldsymbol{T}$, które negatywnie wpływają na jakość wyznaczonej hiperpłaszczyzny i powinny zostać usunięte ze zbioru treningowego). W przypadku dynamicznie adaptacyjnego algorytmu genetycznego (DAGA) zbadano nie tylko nowy schemat adaptacji wielkości zredukowanych zbiorów na podstawie zaproponowanego współczynnika wypełnienia, ale też możliwość zastosowania algorytmu treningu zredukowanego (ang. *reduced* SVM, RSVM) w trakcie optymalizacji genetycznej.

W algorytmie memetycznym (MASVM) zaproponowano rozwiązania algorytmiczne pozwalające na wykorzystanie puli cennych wektorów $\boldsymbol{P}$ do ulepszania istniejących osobników w populacji oraz do tworzenia nowych osobników, zawierających wyłącznie wektory z puli $\boldsymbol{P}$. Podczas analizy czułościowej zbadano jedenaście wariantów algorytmu MASVM, aby dokładnie zweryfikować jak wykorzystanie informacji o potencjalnie istotnych wektorach podczas ewolucji wpływa na jakość końcowych zbiorów treningowych. W algorytmie PCA^{2}MA wprowadzono dodatkowy krok wstępnej analizy pełnego zbioru $\boldsymbol{T}$, mający na celu ekstrakcję wektorów-kandydatów, które są następnie użyte podczas optymalizacji (np. do tworzenia początkowej generacji rozwiązań, a także do kompensacji i tworzenia nowych osobników w czasie ewolucji). Zaproponowany został także bezparametryczny schemat adaptacji wielkości chromosomów w populacji.

Wyniki analizy czułościowej wykazały, że rozwiązania algorytmiczne opisane w przedstawionej rozprawie istotnie polepszają jakość zredukowanych zbiorów treningowych. Wykazano, że algorytmy memetyczne (MASVM oraz PCA^{2}MA) pozwalają na uzyskanie wyraźnie mniejszych zbiorów treningowych (oraz mniejszej liczby wektorów podpierających). Testy Wilcoxona zostały użyte do potwierdzenia statystycznej istotności różnic pomiędzy zredukowanymi zbiorami treningowymi otrzymanymi przy pomocy zaproponowanych algorytmów ewolucyjnych (zweryfikowanie statystycznej istotności otrzymywanych wyników było jednym z drugorzędnych celów przedstawionej rozprawy – **cel ten został osiągnięty**).

1.4.2 Porównanie zaproponowanych algorytmów ewolucyjnych

Zaproponowane algorytmy ewolucyjne zostały wszechstronnie przebadane oraz dokładnie porównane, na podstawie wyników otrzymanych dla trzech zbiorów danych (zbioru benchmarkowego: Adult, sztucznie wygenerowanego: 2D–random$^\alpha$, oraz zawierającego dane rzeczywiste: Skin). W przypadku algorytmu GASVM, wykorzystano różne wielkości osobników w populacji (liczba wektorów w zredukowanym zbiorze nie podlega aktualizacji podczas ewolucji w przypadku tego algorytmu): $t' = 4$ oraz $t' = \bar{t'}$, gdzie $\bar{t'}$ oznacza średnią liczbę wektorów w zbiorach zredukowanych otrzymanych przy użyciu algorytmu PCA^{2}MA. Zbadano także wpływ zaproponowanych rozwiązań algorytmicznych na jakość otrzymywanych rozwiązań – analiza ta była jednym z głównych celów rozprawy.

Tabela 1.1: Wartości funkcji Q otrzymane dla wszystkich zaproponowanych algorytmów ewolucyjnych. Najlepsze wyniki dla każdej wartości parametru q zostały pogrubione, a tło komórek zawierających najmniejsze wartości funkcji Q zostało zaszarzone.

$q \rightarrow$	0.5	0.6	0.7	0.8	0.9	1.0	Średnia
2D–random$^\alpha$							
GASVM(4)	0.8061	0.7674	0.7286	0.6898	0.6511	0.6123	0.7092
GASVM(166)	0.5189	0.6135	0.7081	0.8028	0.8974	0.9920	0.7555
AGA	**0.9891**	**0.9869**	**0.9848**	**0.9726**	0.9805	0.9783	**0.9820**
DAGA	0.6274	0.7017	0.7760	0.8503	0.9246	0.9989	0.8131
MASVM	0.5749	0.6599	0.7450	0.8300	0.9150	**1.0**	0.7875
PCA^{2}MA	0.7192	0.7347	0.8002	0.9457	**0.9812**	0.9967	0.8630
Skin							
GASVM(4)	**0.9864**	**0.9837**	**0.9810**	**0.9783**	0.9756	0.9728	**0.9796**
GASVM(24)	0.6246	0.6978	0.7709	0.8441	0.9173	0.9905	0.8075
AGA	0.8755	0.9002	0.9250	0.9397	0.9644	0.9991	0.9340
DAGA	0.8132	0.8503	0.8873	0.9243	0.9613	0.9983	0.9058
MASVM	0.6921	0.7535	0.8150	0.8764	0.9379	0.9963	0.8452
PCA^{2}MA	0.9153	0.9322	0.9492	0.9731	**0.9831**	**1.0**	0.9588
Adult							
GASVM(4)	0.9722	0.9667	0.9611	0.9556	0.9500	0.9445	0.9583
GASVM(58)	0.5233	0.6142	0.7051	0.7959	0.8868	0.9777	0.7505
AGA	**0.9970**	**0.9964**	**0.9958**	**0.9952**	**0.9946**	0.9940	**0.9955**
DAGA	0.5616	0.6483	0.7351	0.8218	0.9085	0.9953	0.7784
MASVM	0.5710	0.6563	0.7416	0.8270	0.8993	0.9976	0.7821
PCA^{2}MA	0.9725	0.9780	0.9835	0.9890	0.9945	**1.0**	0.9862
Średnia							
GASVM(4)	0.9216	0.9059	0.8902	0.8746	0.8589	0.8432	0.8824
GASVM(t')	0.5556	0.6418	0.7280	0.8143	0.9005	0.9867	0.7712
AGA	**0.9539**	**0.9612**	**0.9685**	0.9692	0.9798	0.9905	**0.9705**
DAGA	0.6674	0.7334	0.7994	0.8655	0.9315	0.9975	0.8325
MASVM	0.6127	0.6899	0.7672	0.8445	0.9174	0.9980	0.8049
PCA^{2}MA	0.8690	0.8817	0.9110	**0.9693**	**0.9863**	**0.9989**	0.9360

Przeanalizowano nie tylko wartości funkcji Q otrzymane dla każdego z zaproponowanych algorytmów (wartości te zostały zebrane w Tabeli 1.1 – wyniki wskazują, że najlepszym algorytmem ewolucyjnym jest PCA^{2}MA dla $q \geq 0.8$; warto zauważyć, że w większości aplikacji $q \approx 1$, ponieważ jakość klasyfikacji jest zwykle „ważniejsza" niż czas działania klasyfikatora, tj. liczba wektorów podpierających), ale także inne własności algorytmów ewolucyjnych (np. liczbę przetworzonych generacji, średnią liczbę wektorów w zredukowanych zbiorach oraz uzyskane współczynniki wypełnienia). Dodatkowo przykładowe zbiory treningowe (wraz z zaznaczonymi wektorami podpierającymi) zostały zwizualizowane dla zbioru 2D–random$^\alpha$, co pozwoliło na zbadanie tego, w jaki sposób kolejne wektory ze zbioru $\boldsymbol{T}$ są dołączane do zredukowanych zbiorów treningowych podczas ewolucji. Wykazano także, że adaptacyjny algorytm memetyczny PCA^{2}MA charakteryzuje się wyraźnie lepszą zbieżnością przeszukiwania w porównaniu z algorytmem MASVM, oraz pozwala na uzyskanie najlepszych zredukowanych zbiorów

treningowych w porównaniu z innymi algorytmami ewolucyjnymi. Testy Wilcoxona potwierdziły statystyczną istotność otrzymanych wyników (jak wspomniano wcześniej, zweryfikowanie statystycznej istotności otrzymywanych wyników było jednym z drugorzędnych celów przedstawionej rozprawy). Uzyskane wyniki **udowadniają pierwszą część (a) pierwszej tezy przedstawionej rozprawy** – zastosowanie technik adaptacji pozwala na uzyskanie wyraźnie lepszych zbiorów zredukowanych w porównaniu z tymi, które zostały otrzymane przy użyciu algorytmów genetycznych, w których takie techniki nie zostały wykorzystane.

Otrzymane wyniki **częściowo potwierdziły prawdziwość drugiej tezy przedstawionej rozprawy** – zastosowanie algorytmów memetycznych, które wykorzystały informacje na temat wektorów w zbiorze treningowym wyekstrahowane przed optymalizacją (w procesie przetwarzania wstępnego), oraz te zdobyte w czasie ewolucji, pozwoliło na uzyskanie wyraźnie lepszych zredukowanych zbiorów treningowych w porównaniu z algorytmami genetycznymi. Wykazano także, że zastosowanie technik adaptacyjnych umożliwia polepszenie i znaczne ułatwienie procesu doboru zbioru treningowego dla klasyfikatora SVM, co stanowi **potwierdzenie prawdziwości drugiej części (b) pierwszej tezy rozprawy**.

1.4.3 Porównanie algorytmów ewolucyjnych z innymi algorytmami znanymi z literatury

Zaproponowane algorytmy ewolucyjne (w szczególności zaś adaptacyjny algorytm memetyczny PCA^{2}MA) zostały porównane z innymi znanymi z literatury algorytmami do doboru zredukowanego zbioru treningowego dla klasyfikatora SVM. Algorytmy znane z literatury, które zostały eksperymentalnie porównane z algorytmami opisanymi w przedstawionej rozprawie należą do różnych grup metod. Zaimplementowano i przebadano następujące algorytmy z literatury: (i) próbkowanie losowe [1], (ii) algorytm treningu zredukowanego (RSVM) [9], (iii) algorytm bazujący na analizie geometrii zbioru treningowego (ang. *sample reduction by data structure analysis*, SR-DSA) [14], oraz (iv) trening klasyfikatora SVM przy użyciu pełnego zbioru $\boldsymbol{T}$. Wyniki otrzymane przy pomocy algorytmów doboru zredukowanego zbioru treningowego zostały także porównane z wynikami (tj. z jakością klasyfikacji oraz z liczbą wektorów podpierających) otrzymanymi w wypadku użycia pełnego zbioru treningowego do treningu klasyfikatora SVM.

Wyniki eksperymentalne wykazały, że zastosowanie algorytmu PCA^{2}MA pozwala na uzyskanie zredukowanych zbiorów treningowych dla klasyfikatora SVM o jakości wyraźnie lepszej w porównaniu z pozostałymi metodami. Zbadane zostały wartości funkcji Q uzyskane dla sześciu przeanalizowanych zbiorów danych: czterech sztucznie wygenerowanych, jednego benchmarkowego oraz jednego zawierającego dane rzeczywiste (uśrednione wartości funkcji Q zostały zaprezentowane w Tabeli 1.2), a także inne charakterystyki uzyskanych zbiorów $\boldsymbol{T'}$ (takie jak ich średnia wielkość, liczba wyznaczonych wektorów podpierających czy

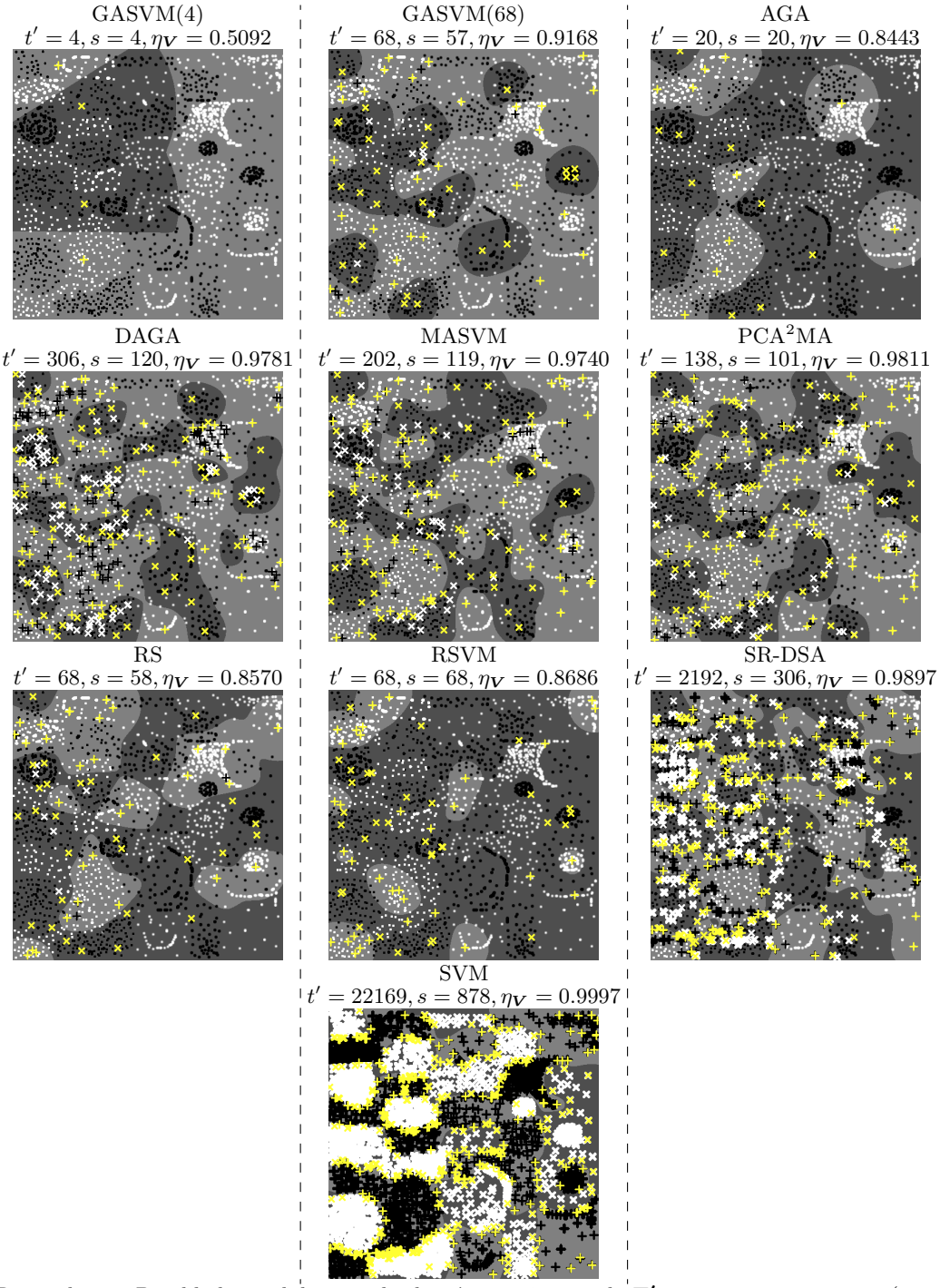
współczynniki wypełnienia). Zbadany został także czas wykonania poszczególnych algorytmów i wykazano, że dla pewnej wielkości zbioru treningowego nie jest możliwy trening klasyfikatora SVM przy użyciu pełnego zbioru $\boldsymbol{T}$ (ze względu na złożoność czasową i pamięciową procesu treningu). Warto zaznaczyć, że czas wykonania zaproponowanych algorytmów ewolucyjnych jest *kontrolowany*, tzn. wykonanie algorytmu może zostać przerwane w odpowiednim momencie (np. gdy zredukowane zbiory treningowe są odpowiedniej jakości). Nie jest to możliwe w przypadku metod analizujących pełny zbiór treningowy (np. SR-DSA).

Otrzymane wyniki ostatecznie potwierdziły **prawdziwość drugiej tezy rozprawy** – zastosowanie algorytmu memetycznego (PCA^{2}MA) pozwoliło na uzyskanie lepszych zredukowanych zbiorów treningowych w porównaniu z tymi, które zostały otrzymane przy użyciu algorytmów genetycznych i innych znanych z literatury.

Tabela 1.2: Uśrednione (dla 6 zbiorów danych) wartości funkcji Q. Najlepsze wyniki dla każdej wartości parametru q zostały pogrubione, a tło komórek zawierających najmniejsze wartości funkcji Q zostało zaszarzone.

$q \rightarrow$	0.5	0.6	0.7	0.8	0.9	1.0	Średnia
SVM	0.5123	0.5763	0.6403	0.7043	0.7683	0.8323	0.6723
RS	0.9232	0.9328	0.9425	0.9521	0.9617	0.9713	0.9473
RSVM	0.8231	0.8492	0.8753	0.9014	0.9275	0.9536	0.8884
SR-DSA	0.7942	0.8312	0.8682	0.9053	0.9419	0.9793	0.8867
PCA^{2}MA	**0.9275**	**0.9385**	**0.9495**	**0.9605**	**0.9715**	**0.9825**	**0.9550**

Wizualizacja otrzymanych zredukowanych zbiorów treningowych wraz z zaznaczonymi wektorami podpierającymi pozwoliła na przeprowadzenie analizy jakościowej oraz zweryfikowanie, które wektory ze zbioru $\boldsymbol{T}$ są oznaczane jako „istotne" przez wszystkie przeanalizowane metody. Warto zauważyć, że wizualizacja otrzymywanych zredukowanych zbiorów treningowych (i wektorów podpierających) była jednym z drugorzędnych celów przedstawionej rozprawy – **cel ten został osiągnięty**. Wykazano, że zredukowane zbiory treningowe otrzymane przy użyciu zaproponowanych algorytmów ewolucyjnych zawierają wektory, które nie zostałyby wybrane do $\boldsymbol{T'}$ przy pomocy metod bazujących wyłącznie na analizie geometrii zbioru $\boldsymbol{T}$. Dołączenie tych wektorów do zredukowanych zbiorów pozwoliło na polepszenie jakości klasyfikacji – wektory te okazały się istotne. Wykazano także, że zastosowanie zaproponowanych metod adaptacji wielkości osobników w populacji wyraźnie wpływa na wielkość i jakość zbiorów $\boldsymbol{T'}$. Algorytm PCA^{2}MA okazał się najlepszym dla redukcji zbiorów treningowych w przypadku wszystkich sztucznie wygenerowanych zbiorów 2D. Wykazano, że użycie pełnego zbioru $\boldsymbol{T}$ do treningu klasyfikatora SVM skutkuje uzyskaniem znacznie większej liczby wektorów podpierających (tj. zwiększeniem czasu klasyfikacji) bez wyraźnej poprawy jakości klasyfikacji. Przykłady zredukowanych zbiorów treningowych otrzymanych dla zbioru 2D–chessboard$^{\alpha}$ zostały zaprezentowane na Rysunku 1.2.



Rysunek 1.2: Przykłady zredukowanych zbiorów treningowych $\boldsymbol{T'}$ wraz z zaznaczonymi (na żółto) wektorami podpierającymi, otrzymanymi dla zbioru 2D–chessboard$^{\alpha}$.

1.5 Podsumowanie i wnioski

W pracy przedstawiono **pięć** algorytmów ewolucyjnych do doboru zredukowanych zbiorów treningowych dla klasyfikatora SVM:

- Algorytm genetyczny (GASVM),
- Adaptacyjny algorytm genetyczny (AGA),
- Dynamicznie adaptacyjny algorytm genetyczny (DAGA),
- Algorytm memetyczny (MASVM), oraz
- Adaptacyjny algorytm memetyczny wykorzystujący dodatkową analizę geometrii danych w zbiorze $\boldsymbol{T}$ (PCA^{2}MA).

Zaproponowane algorytmy zostały wszechstronnie zbadane – zarówno teoretycznie jak i eksperymentalnie – przy użyciu zbiorów danych benchmarkowych, sztucznie wygenerowanych oraz zbioru zawierającego dane rzeczywiste. Szczegółowe badania miały na celu zweryfikowanie istotności wprowadzonych rozwiązań algorytmicznych oraz ich wpływu na jakość uzyskiwanych zredukowanych zbiorów treningowych, a także porównanie zaproponowanych algorytmów ewolucyjnych z innymi technikami doboru zbiorów treningowych znanymi z literatury. W ramach przedstawionej rozprawy rozwinięto nie tylko implementacje powyższych algorytmów ewolucyjnych, ale także programy do automatycznego doboru wartości parametrów funkcji jądrowych, statystycznej analizy otrzymywanych wyników, przetwarzania wsadowego oraz wizualizacji wyników.

Przeprowadzono wszechstronne badania eksperymentalne, które objęły (i) analizę czułościową, (ii) porównanie zaproponowanych algorytmów ewolucyjnych, oraz (iii) porównanie algorytmów ewolucyjnych (w szczególności zaś algorytmu PCA^{2}MA) z innymi algorytmami znanymi z literatury (należącymi do różnych grup metod). W celu łatwiejszego porównania jakości otrzymanych zredukowanych zbiorów treningowych, zdefiniowana została funkcja jakości Q (jak zauważono w Sekcji 1.4, problem doboru zbiorów $\boldsymbol{T'}$ może być interpretowany jako dwukryterialny problem optymalizacyjny).

Wyniki eksperymentalne wykazały, że zredukowane zbiory treningowe otrzymane przy użyciu adaptacyjnych algorytmów genetycznych (AGA oraz DAGA) są wyraźnie lepsze od zbiorów uzyskanych przy użyciu algorytmu genetycznego (GASVM) – jest to **potwierdzenie prawdziwości pierwszej tezy rozprawy**. Istotną cechą zaproponowanych algorytmów ewolucyjnych jest to, że ich czas wykonania może być *kontrolowany*, tj. możliwe jest przerwanie wykonywania algorytmu ewolucyjnego w przypadku, gdy otrzymano zredukowany zbiór treningowy o żądanej jakości. Nie jest to możliwe w przypadku metod bazujących na analizie pełnego zbioru $\boldsymbol{T}$ (np. w przypadku algorytmu SR-DSA). Warto dodać,

że przedstawione algorytmy adaptacyjne nie wymagają przeprowadzenia procesu strojenia parametrów, który jest szczególnie uciążliwy i czasochłonny w przypadku bardzo dużych zbiorów danych – zastosowanie metod adaptacji w algorytmach genetycznych znacznie polepsza proces doboru zbiorów treningowych w porównaniu do algorytmów genetycznych, które nie wykorzystują technik adaptacji, co stanowi **potwierdzenie prawdziwości drugiej części (b) pierwszej tezy rozprawy**. Wykazano także, że zastosowanie algorytmów ewolucyjnych do doboru zredukowanych zbiorów umożliwia usunięcie podczas ewolucji tych wektorów ze zbioru $\boldsymbol{T'}$, które mogą negatywnie wpłynąć na jakość klasyfikacji (a mogły pojawić się w zbiorze zredukowanym np. podczas operacji mutacji). Uzyskane wyniki potwierdziły, że proces generacji początkowej populacji rozwiązań jest istotny i znacznie wpływa na zbieżność algorytmów ewolucyjnych.

Wyniki badań dowiodły, że zredukowane zbiory treningowe otrzymane przy pomocy zaproponowanych algorytmów (zwłaszcza adaptacyjnego algorytmu memetycznego – PCA^{2}MA) są lepsze niż zbiory wyekstrahowane przy pomocy przeanalizowanych metod znanych z literatury oraz innych zaproponowanych algorytmów ewolucyjnych (jest to **potwierdzenie prawdziwości drugiej tezy**). Zaproponowane algorytmy ewolucyjne pozwalają nie tylko na zwiększenie skuteczności klasyfikacji klasyfikatora SVM, którego trening został przeprowadzony przy użyciu otrzymanego zbioru $\boldsymbol{T'}$, ale też na zmniejszenie liczby wektorów podpierających, co z kolei przyspiesza proces klasyfikacji. Warto dodać, że w przypadku zbioru rzeczywistego, zawierającego 4 miliony wektorów, trening klasyfikatora SVM nie był możliwy w przypadku użycia pełnego zbioru $\boldsymbol{T}$, ze względu na jego złożoność czasową oraz pamięciową.

Cele główne zostały uzupełnione celami dodatkowymi – zwizualizowaniem otrzymywanych zredukowanych zbiorów treningowych, oraz analizą statystyczną otrzymywanych wyników. Wizualizacja wyników otrzymanych dla sztucznie wygenerowanych zbiorów 2D pozwoliła na dokładne zweryfikowanie, które z wektorów w zbiorze $\boldsymbol{T}$ są oznaczane jako „istotne" i są umieszczane w zredukowanych zbiorach $\boldsymbol{T'}$. Wykazano, że zastosowanie zaproponowanych algorytmów ewolucyjnych pozwala na dołączenie do zbiorów $\boldsymbol{T'}$ tych wektorów, które zostałyby uznane za zbędne w przypadku użycia innych metod (np. tych, bazujących na analize cech geometryczne pełnego zbioru $\boldsymbol{T}$). Dołączenie tych wektorów umożliwiło dalsze polepszenie jakości klasyfikatorów SVM, których trening został przeprowadzony przy użyciu zbiorów zredukowanych. Statystyczna istotność otrzymywanych wyników (tj. tego, czy różnice pomiędzy zredukowanymi zbiorami treningowymi wyekstrahowanymi przy użyciu przeanalizowanych algorytmów są statystycznie istotne) została zweryfikowana przy użyciu testów Wilcoxona. Zbadano także, czy zaproponowane (w kolejnych algorytmach ewolucyjnych) rozwiązania algorytmiczne istotnie wpływają na otrzymywane wyniki i czy ich zastosowanie jest korzystne, tj. czy prowadzi do uzyskania zredukowanych zbiorów treningowych

o wyższej jakości.

W ramach przedstawionej rozprawy doktorskiej, **udowodnione zostały dwie tezy postawione w Sekcji 1.2**. W celu zweryfikowania prawdziwości postawionych tez, zdefiniowano **siedem celów głównych** oraz **dwa cele drugorzędne – wszystkie cele rozprawy zostały osiągnięte**.

Rozwiązania algorytmiczne rozwinięte w ramach prac nad niniejszą rozprawą znalazły zastosowanie również w algorytmach ewolucyjnych (zwłaszcza memetycznych) do rozwiązywania innych złożonych problemów obliczeniowych [10]. Mogą także zostać wykorzystane w wielu innych domenach, np. w analizie danych medycznych czy optymalizacji kombinatorycznej. Wyniki i algorytmy przedstawione w rozprawie mogą być dalej rozwijane. Potencjalne kierunki dalszych badań obejmują prace nad: dodatkowymi procedurami analizy wstępnej zbioru $\boldsymbol{T}$, analizą zbiorów zaszumionych, automatycznym doborem funkcji jądrowych, oraz skróceniem czasu ewolucji (np. przy użyciu algorytmu równoległego), zastosowaniem zaproponowanych algorytmów do doboru zredukowanych zbiorów treningowych dla innych klasyfikatorów. Ciekawym kierunkiem dalszych badań jest niewątpliwie połączenie algorytmów doboru zredukowanych zbiorów treningowych z algorytmami ekstrakcji i selekcji najbardziej reprezentatywnych cech.

Rozdział 2

Dorobek naukowy (stan na 11 kwietnia 2016 r.)

2.1 Lista publikacji

Wszystkie publikacje zostały podzielone na cztery kategorie – publikacje w czasopismach z listy ministerialnej A (listy filadelfijskiej), publikacje w czasopismach z listy B, publikacje w materiałach konferencyjnych i rozdziałach książek, oraz publikacje zaakceptowane do druku. Dla publikacji w czasopismach z listy JCR (listy filadelfijskiej) podano współczynnik oddziaływania (ang. *impact factor*, IF) oraz liczbę punktów ministerialnych (Ministerstwa Nauki i Szkolnictwa Wyższego – MNiSzW). Sumaryczny współczynnik oddziaływania wynosi **IF=6,176**, a sumaryczna liczba punktów ministerialnych: **272**. Publikacje w materiałach konferencyjnych, które są odnotowane w bazie Web of Science zostały podkreślone (dla każdej z tych publikacji, liczba punktów MNiSzW wynosi 10).

Publikacje w czasopismach z listy A (lista filadelfijska): 5 publikacji

1. **Nalepa J.**, Kawulok M., "Adaptive memetic algorithm enhanced with data geometry analysis to select training data for SVMs", Neurocomputing, s. 1-20, DOI: 10.1016/j.neucom.2015.12.046, Elsevier, 2016.
   IF=2,083, Liczba pkt. MNiSzW: 30

2. **Nalepa J.**, Błocho M., "Adaptive memetic algorithm for minimizing distance in the vehicle routing problem with time windows", Soft Computing, s. 1-19, DOI: 10.1007/s00500-015-1642-4, Springer, 2015.
   IF=1,271, Liczba pkt. MNiSzW: 25

3. **Nalepa J.**, Błocho M., "Co-operation in the parallel memetic algorithm", International Journal of Parallel Programming, Vol. 43, Issue 5, s. 812-839, DOI: 10.1007/s10766-014-0343-4, Springer, 2015.
IF=0,491, Liczba pkt. MNiSzW: 15

4. Kawulok M., Kawulok J., **Nalepa J.**, Smołka B., "Self-adaptive algorithm for segmenting skin regions", EURASIP Journal on Advances in Signal Processing, Vol. 2014, Nr 170, s. 1-22, DOI: 10.1186/1687-6180-2014-170, 2014.
IF=0,78, Liczba pkt. MNiSzW: 25

5. Kawulok M., Kawulok J., **Nalepa J.**, "Spatial-based skin detection using discriminative skin-presence features", Pattern Recognition Letters, Vol. 41, s. 3–13, DOI: 10.1016/j.patrec.2013.08.028, 2014.
IF=1,551, Liczba pkt. MNiSzW: 25

Publikacje w czasopismach z listy B: 2 publikacje

6. **Nalepa J.**, Czech Z. J., "Adaptive threads co-operation schemes in a parallel heuristic algorithm for the vehicle routing problem with time windows", Theoretical and Applied Informatics, Vol. 24, Nr 3, s. 191-203, DOI: 10.2478/v10179-012-0012-5, 2012.
Liczba pkt. MNiSzW: 4

7. **Nalepa J.**, Czech Z. J., "A parallel heuristic algorithm to solve the vehicle routing problem with time windows", Studia Informatica, Vol. 33, Nr 104, s. 91-106, 2012.
Liczba pkt. MNiSzW: 4

Publikacje w materiałach konferencyjnych i rozdziałach książek: 25 publikacji

8. **Nalepa J.**, Błocho M., "Enhanced guided ejection search for the pickup and delivery problem with time windows", 8th Asian Conference on Intelligent Information and Database Systems, ACIIDS 2016, Nguyen N. T., Trawiński B., Fujita H., Hong T. (ed.), Intelligent Information and Database Systems, Vol. 9621, Lecture Notes in Computer Science, s. 388-398, DOI: 10.1007/978-3-662-49381-6_37, Springer, 2016.

9. Ćwięk M., **Nalepa J.**, Dublański M., "How to generate benchmarks for rich routing problems?", 8th Asian Conference on Intelligent Information and Database Systems, ACIIDS 2016, Nguyen N. T., Trawiński B., Fujita H., Hong T. (ed.), Intelligent Information and Database Systems, Vol. 9621,

Lecture Notes in Computer Science, s. 399-409, DOI: 10.1007/978-3-662-49381-6_38, Springer, 2016.

10. Dworak K., **Nalepa J.**, Boryczka U., Kawulok M., "Cryptanalysis of SDES using genetic and memetic algorithms", 8th Asian Conference on Intelligent Information and Database Systems, ACIIDS 2016, Król D., Madeyski L., Nguyen N. T. (ed.), Recent Developments in Intelligent Information and Database Systems, Vol. 642, Studies in Computational Intelligence, s. 3-14, DOI: 10.1007/978-3-319-31277-4_1, Springer, 2016.

11. **Nalepa J.**, Błocho M., "A parallel algorithm with the search space partition for the pickup and delivery with time windows", P2P, Parallel, Grid, Cloud and Internet Computing, 10th IEEE International Conference on, s. 92-99, DOI: 10.1109/3PGCIC.2015.12, 2015.

12. **Nalepa J.**, Ćwięk M., Kawulok M., "Adaptive memetic algorithm for the job shop scheduling problem", Neural Networks (IEEE IJCNN), 2015 International Joint Conference on, s. 1-8, DOI: 10.1109/IJCNN.2015.7280409, 2015.

13. **Nalepa J.**, Szymanek J., Kawulok M., "Real-time people counting from depth images", Beyond Databases, Architectures, and Structures, 11th International Conference, BDAS 2015, Kozielski S., Mrozek D., Kasprowski P., Małysiak-Mrozek B. & Kostrzewa D. (ed.), Communications in Computer and Information Science, Vol. 521, s. 387-397, DOI: 10.1007/978-3-319-18422-7_34, Springer International Publishing, 2015.

14. Błocho M., **Nalepa J.**, "Impact of parallel memetic algorithm parameters on its efficacy", Beyond Databases, Architectures, and Structures, 11th International Conference, BDAS 2015, Kozielski S., Mrozek D., Kasprowski P., Małysiak-Mrozek B. & Kostrzewa D. (ed.), Communications in Computer and Information Science, Vol. 521, s. 299-308, DOI: 10.1007/978-3-319-18422-7_27, Springer International Publishing, 2015.

15. Błocho M., **Nalepa J.**, "A parallel algorithm for minimizing the fleet size in the pickup and delivery problem with time windows", Proceedings of the 22nd European MPI Users' Group, EuroMPI '15, s. 1-2, DOI: 10.1145/2802658.2802673, ACM, 2015.

16. **Nalepa J.**, Kawulok M., "A memetic algorithm to select training data for support vector machines", Proceedings of the 2014 Annual Conference on Genetic and Evolutionary Computation, GECCO 2014, s. 573-580, DOI: 10.1145/2576768.2598370, ACM, 2014.

17. **Nalepa J.**, Kawulok M., "Adaptive genetic algorithm to select training data for support vector machines", Applications of Evolutionary Computation, 17th European Conference, EvoApplications 2014, Esparcia-Alcazar A. I. & Mora A. M., (ed.), Lecture Notes in Computer Science, Vol. 8602, s. 514-525, DOI: 10.1007/978-3-662-45523-4_42, Springer Berlin Heidelberg, 2014.

18. **Nalepa J.**, Błocho M., Czech Z. J., "Co-operation schemes for the parallel memetic algorithm", Parallel Processing and Applied Mathematics, 10th International Conference, PPAM 2013, Wyrzykowski R., Dongarra J., Karczewski K. & Waśniewski J., (ed.), Lecture Notes in Computer Science, Vol. 8384, s. 191-201, DOI: 10.1007/978-3-642-55224-3_19, Springer Berlin Heidelberg, 2014.

19. **Nalepa J.**, Szymanek J., Hayball M. P., Brown S. J., Ganeshan B., Miles K. A., "Texture analysis for identifying heterogeneity in medical images", Computer Vision and Graphics, International Conference, ICCVG 2014, Chmielewski L. J., Kozera R., Bok-Suk S., & Wojciechowski K. (ed.), Lecture Notes in Computer Science, Vol. 8671, s. 446-453, DOI: 10.1007/978-3-319-11331-9_53, Springer International Publishing, 2014.

20. Kawulok M., **Nalepa J.**, "Dynamically adaptive genetic algorithm to select training data for SVMs", Advances in Artificial Intelligence – IBERAMIA 2014, 14th Ibero-American Conference on AI, Bazzan A. L. C. & Pichara K., (ed.), Lecture Notes in Computer Science, Vol. 8864, s. 242-254, DOI: 10.1007/978-3-319-12027-0_20, Springer International Publishing, 2014.

21. **Nalepa J.**, "Adaptive memetic algorithm for the vehicle routing problem with time windows", Proceedings of the Companion Publication of the 2014 Annual Conference on Genetic and Evolutionary Computation, GECCO 2014, s. 1467-1468, DOI: 10.1145/2598394.2602273, ACM, 2014.

22. Ćwięk M., **Nalepa J.**, "A fast genetic algorithm for the flexible job shop scheduling problem", Proceedings of the Companion Publication of the 2014 Annual Conference on Genetic and Evolutionary Computation, GECCO 2014, s. 1449–1450, DOI: 10.1145/2598394.2602280, ACM, 2014.

23. Kawulok M., **Nalepa J.**, "Hand pose estimation using support vector machines with evolutionary training", Systems, Signals and Image Processing (IWSSIP), 2014 IEEE International Conference on, s. 87-90, 2014.

24. **Nalepa J.**, Kawulok M., "Fast and accurate hand shape classification", Beyond Databases, Architectures, and Structures, 10th International Conference, BDAS 2014, Kozielski S., Mrozek D., Kasprowski P., Małysiak-

Mrozek B. & Kostrzewa D. (ed.), Communications in Computer and Information Science, Vol. 424, s. 364-373, DOI: 10.1007/978-3-319-06932-6_35, Springer International Publishing, 2014.

25. **Nalepa J.**, Grzejszczak T., Kawulok M., "Wrist localization in color images for hand gesture recognition", Man-Machine Interactions 3, Gruca A., Czachórski T., & Kozielski S.(ed.), Advances in Intelligent Systems and Computing, Vol. 242, s. 79-86, DOI: 10.1007/978-3-319-02309-0_8, Springer International Publishing, 2014.

26. Kawulok M., Kawulok J., **Nalepa J.**, Smołka B., "Self-adaptive skin segmentation in color images", Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications, 19th Iberoamerican Congress, CIARP 2014, Bayro-Corrochano E. & Hancock E.(ed.), Lecture Notes in Computer Science, Vol. 8827, s. 96-103, DOI: 10.1007/978-3-319-12568-8_12, Springer International Publishing, 2014.

27. Kawulok M., **Nalepa J.**, Kawulok J., "Skin detection and segmentation in color images", Advances in Low-Level Color Image Processing, Celebi M. E. & Smołka B.(ed.), Lecture Notes in Computational Vision and Biomechanics, Vol. 11, s. 329–366, DOI: 10.1007/978-94-007-7584-8_11, Springer Netherlands, 2014.
Liczba pkt. MNiSzW: 4

28. **Nalepa J.**, Czech Z. J., "New selection schemes in a memetic algorithm for the vehicle routing problem with time windows", Adaptive and Natural Computing Algorithms, 11th International Conference, ICANNGA 2013, Tomassini M., Antonioni A., Daolio F., & Buesser P., (ed.), Lecture Notes in Computer Science, Vol. 7824, s. 396-405, DOI: 10.1007/978-3-642-37213-1_41, Springer Berlin Heidelberg, 2013.

29. **Nalepa J.**, Kawulok M., "Parallel hand shape classification", Multimedia (ISM), 2013 IEEE International Symposium on, s. 401–402, DOI: 10.1109/ISM.2013.76, 2013.

30. Grzejszczak T., **Nalepa J.**, Kawulok M., "Real-time wrist localization in hand silhouettes", Proceedings of the 8th International Conference on Computer Recognition Systems CORES 2013, Advances in Intelligent Systems and Computing, Burduk R., Kurzyński M., Woźniak M., Żołnierek A. (ed.), Vol. 226, s. 439-449, DOI: 10.1007/978-3-319-00969-8_43, Springer International Publishing, 2013.

31. Kawulok M., Kawulok J., **Nalepa J.**, Papież M., "Skin detection using spatial analysis with adaptive seed", Image Processing (ICIP), 2013 20th

IEEE International Conference on, s. 3720-3724, DOI: 10.1109/ICIP.2013.6738767, 2013.

32. Kawulok M., **Nalepa J.**, "Support vector machines training data selection using a genetic algorithm", Structural, Syntactic, and Statistical Pattern Recognition, Joint IAPR International Workshop, SSPR & SPR 2012, Gimel'farb G., Hancock E., Imiya A., Kuijper A., Kudo M., Omachi S., Windeatt T., & Yamada K. (ed.), Lecture Notes in Computer Science, Vol. 7626, s. 557-565, DOI: 10.1007/978-3-642-34166-3_61, Springer Berlin Heidelberg, 2012.

Publikacje zaakceptowane do druku: 3 publikacje

33. Kawulok M., Papież M., **Nalepa J.**, "Manifold learning for hand pose recognition: evaluation framework", Beyond Databases, Architectures, and Structures, 11th International Conference, BDAS 2016, Springer International Publishing, 2016.

34. **Nalepa J.**, Simiński K., Kawulok M., "Towards parameter-less support vector machines", Proceedings of the 3rd IAPR Asian Conference on Pattern Recognition, ACPR 2015, s. 1-5, 2015.

35. Kawulok M., **Nalepa J.**, "Towards robust SVM training from weakly labeled large data sets", Proceedings of the 3rd IAPR Asian Conference on Pattern Recognition, ACPR 2015, s. 1-5, 2015.

2.2 Dane bibliometryczne

Dane bibliometryczne dotyczące moich publikacji, zawarte w bazach Web of Science, Scopus, Google Scholar oraz DBLP zostały zaprezentowane w Tabeli 2.1.

Tabela 2.1: Dane bibliometryczne według różnych baz danych.

	Liczba	Liczba cytowań		Indeks
	publikacji	wszystkie	bez autocytowań	Hirscha
Web of Science	17	56	22	5
Scopus	27	73	27	5
Google Scholar	34	163	—	8
DBLP	27	—	—	—

2.3 Najważniejsze projekty badawcze

Poniżej przedstawiam listę projektów badawczych, w których brałem lub biorę udział. Dla każdego projektu wyróżniłem daty jego trwania, pełnioną przeze mnie rolę oraz kwotę i źródło dofinansowania.

02/2016 – 12/2018 **Enhancing the diagnostic efficiency of dynamic contrast-enhanced imaging in personalised oncology by extracting new and improved biomarkers**
Grant Innomed (POIR/01.02.00-24-32/15)
Instytucja przyznająca: Narodowe Centrum Badań i Rozwoju
Kwota dofinansowania: 9 144 266,80 zł
Charakter udziału: Ekspert do spraw algorytmów ewolucyjnych i równoległych oraz uczenia maszynowego

04/2014 – 04/2017 **A parallel memetic algorithm for solving complex optimization problems**
Grant Preludium (DEC-2013/09/N/ST6/03461)
Instytucja przyznająca: Narodowe Centrum Nauki
Kwota dofinansowania: 148 512 zł
Charakter udziału: Kierownik projektu, główny wykonawca

09/2015 – 11/2015 **Odwzorowanie algorytmów optymalizacji kombinatorycznej na przykładzie algorytmu wyznaczania tras na architekturze masywnie wielordzeniowej Intel Xeon Phi**
Instytucja przyznająca: Miclab (`http://miclab.pl/`) – laboratorium pilotażowe systemów masywnie wielordzeniowych (Projekt współfinansowany ze środków Unii Europejskiej – POIG.02.03.00.24-093/13)
Kwota dofinansowania: 7 500 zł
Charakter udziału: Kierownik projektu

06/2013 – 06/2015 **Evolutionary methods for support vector machines training set optimization**
Grant Iuventus Plus (IP2012 026372)
Instytucja przyznająca: Ministerstwo Nauki i Szkolnictwa Wyższego
Kwota dofinansowania: 191 750 zł
Charakter udziału: Główny wykonawca
Kierownik projektu: dr inż. Michał Kawulok

04/2012 – 04/2014 **Hand detection and pose estimation for creating human-computer interaction**
Grant Iuventus Plus (IP2011 023071)
Instytucja przyznająca: Ministerstwo Nauki i Szkolnictwa Wyższego
Kwota dofinansowania: 258 125 zł
Charakter udziału: Wykonawca
Kierownik projektu: dr inż. Michał Kawulok

Oprócz powyższych projektów (finansowanych ze źródeł zewnętrznych), prowadzę badania w ramach Badań Kierunkowych Młodych Naukowców (BKM) w Politechnice Śląskiej (od roku 2012 do teraz). Tematyka projektów finansowanych w ramach BKM obejmuje:

- Rozwój równoległych algorytmów memetycznych dla rozwiązywania problemu trasowania pojazdów z oknami czasowymi,
- Rozwój algorytmów ewolucyjnych dla rozwiązywania złożonych problemów obliczeniowych.

2.4 Przeprowadzone recenzje

Dotychczas recenzowałem artykuły zgłoszone do publikacji w następujących czasopismach z listy filadelfijskiej:

- IEEE Transactions on Evolutionary Computation, IF=3,654 (2 recenzje),
- IEEE Transactions on Industrial Informatics, IF=8,785 (1 recenzja),
- Computers & Operations Research, IF=1,861 (2 recenzje),
- Soft Computing, IF=1,271 (5 recenzji),
- KSII Transactions on Internet and Information Systems, IF=0,561 (1 recenzja).

Recenzowałem także artykuły zgłoszone na konferencje międzynarodowe:

- International Conference on Man-Machine Interactions, ICMMI 2015 (Kocierz, Polska) (2 recenzje),
- The 8th International Conference on Neural Network and Artificial Intelligence (ICNNAI 2014) (Brześć, Białoruś) (3 recenzje); również jako członek Komitetu Organizacyjnego.

2.5 Najważniejsze wyróżnienia i nagrody

Do najważniejszych wyróżnień i nagród zaliczam:

- Zaproszenie do przesłania rozszerzonej wersji artykułu przedstawionego na konferencji CIARP 2014 w Puerto Vallarta, w Meksyku (zaproszono autorów 15 najlepszych artykułów) do publikacji w czasopiśmie Journal of Intelligent Data Analysis (IF=0,606). Artykuł jest obecnie w recenzji.
- Zaproszenie do przesłania rozszerzonej wersji artykułu przedstawionego na konferencji ACIIDS 2016 w Da Nang, w Wietnamie (zaproszono autorów najlepszych artykułów) do publikacji w czasopiśmie Journal of Intelligent & Fuzzy Systems (IF=1,812).
- Stypendium w ramach Funduszu Stypendialno-Stażowego „Przedsiębiorczy naukowiec" (Technopark Gliwice, projekt współfinansowany ze środków Unii Europejskiej), 2013-2015.
- *Best presentation award*, za pracę zaprezentowaną na konferencji IEEE ISM International Symposium on Multimedia (PhD Workshop), Anaheim, USA, 2013.
- II miejsce w 28. Konkursie na Najlepszą Pracę Magisterską z Informatyki (`http://www.pti.org.pl/`) organizowanym przez Polskie Towarzystwo Informatyczne, edycja 2010/2011.
- Stypendium (w Politechnice Śląskiej) dla najlepszych doktorantów (2011, 2012, 2013, 2014, 2015).
- Stypendium (w Politechnice Śląskiej) dla najlepszych studentów (2006, 2007, 2008, 2009, 2010).

Bibliografia

[1] J. L. Balcázar, Y. Dai, and O. Watanabe. A random sampling technique for training support vector machines. In *Proceedings of the International Conference on Algorithmic Learning Theory*, s. 119–134. Springer, 2001.

[2] D. Cheng, J. Wang, X. Wei, and Y. Gong. Training mixture of weighted SVM for object detection using EM algorithm. *Neurocomputing*, 149, Part B:473 – 482, 2015.

[3] C. Cortes and V. Vapnik. Support-Vector Networks. *Machine Learning*, 20(3):273–297, 1995.

[4] B. Cyganek, B. Krawczyk, and M. Woźniak. Multidimensional data classification with chordal distance based kernel and support vector machines. *Engineering Applications of Artificial Intelligence*, 46, Part A:10 – 22, 2015.

[5] P. Insom, C. Cao, P. Boonsrimuang, D. Liu, A. Saokarn, P. Yomwan, and Y. Xu. A support vector machine-based particle filter method for improved flooding classification. *Geoscience and Remote Sensing Letters, IEEE*, 12(9):1943–1947, 2015.

[6] M. Kawulok. Genetic algorithms for classifiers' training sets optimization applied to human face recognition. *Journal of Medical Informatics & Technologies*, 11:135–143, 2007.

[7] L. Khedher, J. Ramirez, J. Gorriz, A. Brahim, and F. Segovia. Early diagnosis of Alzheimer's disease based on partial least squares, principal component analysis and support vector machine using segmented MRI images. *Neurocomputing*, 151, Part 1:139 – 150, 2015.

[8] K. Kim and D. Lee. Inductive manifold learning using structured support vector machine. *Pattern Recognition*, 47(1):470 – 479, 2014.

[9] Y.-J. Lee and S.-Y. Huang. Reduced support vector machines: A statistical theory. *Neural Networks, IEEE Transactions on*, 18(1):1–13, 2007.

[10] J. Nalepa and M. Błocho. Adaptive memetic algorithm for minimizing distance in the vehicle routing problem with time windows. *Soft Computing*, s. 1–19, 2015. (doi:10.1007/s00500-015-1642-4).

[11] J. Nalepa and M. Kawulok. A memetic algorithm to select training data for support vector machines. In *Proceedings of the 2014 Conference on Genetic and Evolutionary Computation*, GECCO '14, s. 573–580. ACM, 2014.

[12] J. Nalepa and M. Kawulok. Adaptive memetic algorithm enhanced with data geometry analysis to select training data for SVMs. *Neurocomputing*, 185:113 – 132, 2016.

[13] I. Rodriguez-Lujan, C. S. Cruz, and R. Huerta. Hierarchical linear support vector machine. *Pattern Recognition*, 45(12):4414 – 4427, 2012.

[14] D. Wang and L. Shi. Selecting valuable training samples for SVMs via data structure analysis. *Neurocomputing*, 71:2772–2781, 2008.

[15] J. Yan, J. Li, and X. Gao. Chinese text location under complex background using Gabor filter and SVM. *Neurocomputing*, 74(17):2998 – 3008, 2011.

[16] G. Zhai, J. Chen, S. Wang, K. Li, and L. Zhang. Material identification of loose particles in sealed electronic devices using PCA and SVM. *Neurocomputing*, 148:222 – 228, 2015.

[17] B. Zhang, J. Yin, S. Wang, and X. Yan. Research on virus detection technique based on ensemble neural network and SVM. *Neurocomputing*, 137:24 – 33, 2014.