Akademia Górniczo–Hutnicza
im. Stanisława Staszica

Wydział Elektrotechniki, Automatyki,
Informatyki i Elektroniki

Katedra Informatyki

Autoreferat rozprawy doktorskiej

Visual Clustering Methods for Pattern Recognition in Biomedical Data

Metody klasteryzacji wizualizacyjnej dla potrzeb rozpoznawania wzorców w danych biomedycznych

mgr inż. Marcin Kurdziel

Promotor:
Dr hab. inż. Krzysztof Boryczko, prof. n. AGH

Kraków, 2010

Wprowadzenie

Praca doktorska omawiana w niniejszym autoreferacie poświęcona jest zagadnieniu detekcji skupisk (ang. *cluster recognition; cluster analysis*) w zbiorach danych zawierających szum, ze szczególnym uwzględnieniem danych pozyskiwanych w technikach pomiarowych biologii molekularnej oraz nowoczesnych metodach diagnostyki medycznej. Szum w zbiorach danych może wiązać się z występowaniem w nich obserwacji nie odnoszących się bezpośrednio do przedmiotu badań (tzw. *szum tła* – ang. *background noise*) lub też z zaburzeniem poszczególnych obserwacji wynikającym z niedokładności metody pomiarowej (tzw. *szum addytywny* – ang. *additive noise*). Za przykład posłużyć może opracowana pod koniec lat dziewięćdziesiątych XX wieku *technika mikromacierzy DNA*. Pozwala ona na równoczesny pomiar poziomów ekspresji wielu tysięcy genów. Pozyskiwane poziomy ekspresji są jednak obarczone znaczącymi błędami pomiarowymi, które próbuje się redukować specjalnie w tym celu opracowanymi metodami normalizacji danych mikromacierzowych. Co więcej, w typowym eksperymencie z wykorzystaniem mikromacierzy jedynie niektóre z genów, których poziom ekspresji jest oznaczany, okazują się być potencjalnie powiązane z badanym procesem biologicznym. Technika mikromacierzy nie jest jedyną metodą biologii molekularnej generującą znaczne ilości danych. W wielu innych obszarach nauk biomedycznych zaawansowane techniki eksperymentalne umożliwiły pozyskanie na tyle obszernych zbiorów danych, by konieczne stało się opracowanie metod pozwalających na wskazanie w dostępnych danych obserwacji potencjalnie istotnych dla wybranego przedmiotu badań.

W klasycznych algorytmach analizy skupisk, pojedyncze skupisko rozumiane jest jako grupa obserwacji podobnych w zadanej mierze odległości, wyodrębniona przestrzennie od innych obserwacji w zbiorze danych. Zakłada się przy tym, iż każda obserwacja w zbiorze danych przynależy do jednego skupiska lub, w przypadku metod *rozmytej analizy skupisk* (ang. *fuzzy clustering methods*), ma przypisany zestaw współczynników przynależności do poszczególnych skupisk. Tak postawiona definicja skupiska jest jednak nieprzydatna w przypadku zbiorów danych obarczonych znaczącym poziomem szumu. Obecność szumu może bowiem spowodować brak wyraźnego, przestrzennego wyodrębnienia jakiejkolwiek grupy obserwacji. Co więcej, obserwacji będących manifestacją szumu tła w zbiorze danych nie należy traktować jako elementów skupisk. W przypadku zbiorów danych obarczonych szumem konieczne jest więc postawienie nowej definicji skupiska – w takim przypadku skupiskiem będzie część zbioru danych charakteryzująca się lokalną gęstością obserwacji wyższą od gęstości obserwacji w otaczającym ją szumie tła.

Począwszy od lat 90-tych algorytmy analizy skupisk oparte na kryterium lokalnej gęstości obserwacji są przedmiotem intensywnych prac badawczych. W publikowanych pracach skuteczność proponowanych algorytmów oceniana jest zazwyczaj na podstawie rezultatów analizy skupisk dla kilku testowych zbiorów danych. Brak jest natomiast ilościowej oceny skuteczności proponowanych metod w zależności od poziomu szumu w analizowanym zbiorze danych. Co więcej, doświadczenia autora referowanej rozprawy doktorskiej wskazują, iż nawet szeroko cytowane metody detekcji skupisk na podstawie kryterium gęstości obserwacji są w istocie stosunkowo mało odporne na narastający poziom szumu w danych. Niska skuteczność okazuje się być wynikiem wrażliwości stosowanych w tych algorytmach technik estymacji gęstości na wzrost amplitudy szumu addytywnego lub też wzrost ilości obserwacji

w obrębie szumu tła. W praktycznych zastosowaniach wspomnianych algorytmów problematyczny jest także dobór odpowiednich wartości parametrów, które pozwalałyby usunąć ze zbioru danych szum tła bez znaczącego uszczerbku dla występującej w nim struktury skupisk.

Teza i cele rozprawy

W referowanej pracy postawiłem następującą tezę: *rozkład gęstości w zbiorze obserwacji można oszacować za pomocą wielokrotnej wymiany komunikatów wzdłuż krawędzi grafu k–najbliższych sąsiadów, w którym wierzchołki reprezentują obserwacje. Treścią komunikatów przesyłanych pomiędzy wierzchołkami są poprawki do oszacowań wartości gęstości, wyznaczone na podstawie pozycji wierzchołków w posortowanych listach sąsiedztwa grafu. Proponowana metoda estymacji gęstości może być stosowana zarówno w przypadku obserwacji będących wielowymiarowymi wektorami cech jak i zbiorów danych opisanych za pomocą macierzy wartości miary niepodobieństwa obserwacji. Proponowany estymator umożliwia skonstruowanie algorytmu analizy skupisk w danych obarczonych szumem addytywnym oraz szumem tła, charakteryzującego się większą dokładnością niż klasyczne algorytmy rozpoznające skupiska na podstawie kryterium gęstości obserwacji. W połączeniu z metodą wizualizacji zbiorów danych, opartą na technice skalowania wielowymiarowego poprzez minimalizację funkcji naprężenia wizualizacji, proponowany algorytm analizy skupisk rozpoznaje w danych biomedycznych grupy obserwacji zgodne z naturą badanych obiektów lub procesów.*

Celem rozprawy jest zaproponowanie odpornej na szum, zarówno addytywny jak i tła, metody estymacji rozkładów gęstości w zbiorach danych i opracowanie na jej podstawie algorytmu analizy skupisk dla danych o wysokim poziomie szumu. W tym zakresie nacisk położony jest zarówno na opracowanie nowych algorytmów jak i na ilościową ocenę ich odporności na szum i odniesienie otrzymanych wyników do analogicznej oceny metod zaproponowanych w literaturze. Głównym obszarem zastosowań proponowanych algorytmów jest analiza danych pozyskiwanych w biologii molekularnej i medycynie. W wielu wypadkach dane te nie mają charakteru wielowymiarowych wektorów cech lecz składają się na nie obiekty o bardziej złożonej strukturze, które można jedynie porównywać za pomocą odpowiednio w tym celu zaprojektowanych miar niepodobieństwa. Realizacja postawionego celu wymaga więc opracowania algorytmów przydatnych nie tylko dla danych wektorowych lecz także danych opisanych przez macierze wartości miary niepodobieństwa dla par obserwacji.

Praktyczne zastosowania metod detekcji skupisk w danych obarczonych szumem wiążą się z koniecznością doboru wartości parametrów, które mogą w istotny sposób wpływać na końcowy rezultat analizy. W szczególności, algorytmy te wymagają wskazania jaki odsetek obserwacji w zbiorze danych wydaje się być częścią szumu tła lub alternatywnie, jaka jest graniczna wartość gęstości obserwacji będących manifestacją szumu. W przypadku szumu o gęstości niewielkiej w porównaniu do gęstości skupisk, maksymalną gęstość obserwacji w szumie tła można ustalić na podstawie histogramu wartości gęstości oszacowanych dla analizowanego zbioru danych. W trudniejszych problemach histogram gęstości może jednak nie dać jednoznacznego oszacowania dla tego parametru. Parametry algorytmów analizy skupisk mogą także wpływać na „rozdzielczość" konstruowanych struktur skupisk,

prowadząc, w zależności od wybranych wartości, do uzyskania niewielkiej liczby dużych skupisk lub też większej liczby skupisk mniejszych. Powyższe zagadnienia wskazują kolejny cel rozprawy, którym jest opracowanie metody wizualizacji zbiorów danych, pozwalającej na dobór odpowiednich wartości parametrów dla proponowanych algorytmów estymacji gęstości i analizy skupisk oraz umożliwiającej weryfikację zgodności konstruowanych struktur skupisk z rozkładem obserwacji w zbiorze danych. Z powodów przedstawionych w poprzednim akapicie, metoda ta powinna być przydatna zarówno dla danych wektorowych jak i danych opisanych przez wartości miary niepodobieństwa dla par obserwacji.

Ostatnim celem niniejszej rozprawy jest weryfikacja skuteczności zaproponowanych algorytmów w rzeczywistych zagadnieniach analizy danych biomedycznych. W tym miejscu warto zwrócić uwagę na różnorodność typów danych, z którymi mamy do czynienia w biologii i medycynie. W zagadnieniach związanych z obrazowaniem medycznym konstruowane są zbiory danych wektorowych opisujące struktury widoczne na pozyskiwanych zdjęciach. Mogą to być np. statystyczne cechy rozkładów intensywności pikseli na fragmentach zdjęć mammograficznych. W ramach eksperymentów mikromacierzowych pozyskiwanie są dane w postaci serii czasowych opisujących przebieg poziomów ekspresji genów w trakcie różnorodnych zjawisk biologicznych lub też procesów chorobowych. W biologii molekularnej znaczącą rolę odgrywa analiza sekwencji polimerów biologicznych, takich jak cząsteczki DNA i RNA lub łańcuchy białkowe. W tym przypadku analizowanymi danymi są napisy reprezentujące badane sekwencje. Z jeszcze innym typem danych mamy do czynienia w przypadku badań strukturalnych w biologii molekularnej. Tutaj obserwacją w zbiorze danych jest opis trójwymiarowej struktury biocząsteczki, np. białka, w postaci współrzędnych przestrzennych składających się na nią atomów. Demonstracja przydatności proponowanych algorytmów w problemach analizy danych związanych z współczesną biologią i medycyną musi uwzględniać wszystkie wymienione tu podstawowe typy danych biomedycznych.

Zawartość rozprawy

Główną część rozprawy otwiera rozdział prezentujący formalnie problem detekcji skupisk w zbiorach danych oraz wprowadzający definicję dwóch podstawowych kategorii szumu w danych, tj. szumu addytywnego i szumu tła. Rozdział ten wskazuje także dostępne w literaturze podstawowe wyniki teoretyczne dotyczące NP–zupełności problemu analizy skupisk. W tym kontekście podkreślona jest konieczność poszukiwania algorytmów konstruujących suboptymalne struktury skupisk przy akceptowalnej, z punktu widzenia dostępnej mocy obliczeniowej oraz rozmiarów analizowanych zbiorów danych, złożoności obliczeniowej. Kolejny rozdział prezentuje opublikowane w literaturze najważniejsze klasyczne algorytmy analizy skupisk oraz wszystkie znaczące rezultaty dotyczące analizy skupisk dla danych obarczonych szumem. Dalsza część przeglądu literatury poświęcona jest metodom wizualizacji zbiorów danych. W tym zakresie omawiane są zarówno klasyczne algorytmy wizualizacji jak i techniki odwzorowywania zbiorów danych, w których obserwacje leżą na wielowymiarowych powierzchniach (ang. *manifold mapping techniques*). Rozdziały te uwzględniają dostępne w literaturze rezultaty ocen skuteczności prezentowanych metod oraz omawiają potencjalny zakres ich zastosowań i główne ograniczenia z perspektywy celów niniejszej pracy.

Kolejne rozdziały rozprawy prezentuję wyniki moich prac badawczych. W pierwszej kolejności proponuję nowy algorytm estymacji gęstości. W szczególności, prezentuję metodę oszacowania poziomów gęstości na podstawie pozycji obserwacji w listach sąsiedztwa grafu k–najbliższych sąsiadów zbudowanego nad zbiorem obserwacji. Następnie proponuję nowy algorytm analizy skupisk wykorzystujący zaproponowaną metodę estymacji gęstości. Algorytm ten na podstawie oszacowanych poziomów gęstości poszukuje tzw. *obserwacji reprezentatywnych*, wokół których konstruowane są poszczególne skupiska. W dalszej części rozprawy prezentuję metodę wizualizacji danych ułatwiającą praktyczne stosowanie zaproponowanych algorytmów estymacji gęstości i analizy skupisk. Proponowana metoda wizualizacji jest rozwinięciem opisanej w literaturze tzw. *cząsteczkowej metody skalowania wielowymiarowego*. Proponuję także heurystyczną technikę przyspieszania procesu wizualizacji danych omawianym algorytmem.

Następny rozdział pracy doktorskiej prezentuje wyniki, które uzyskałem w ramach eksperymentalnej oceny skuteczności zaproponowanych metod. W pierwszej części rozdziału ilustruję działanie metody estymacji gęstości i algorytmu analizy skupisk na przykładzie kilku prostych zbiorów danych. Następnie prezentuję ilościowe wyniki dotyczące zbieżności zaproponowanej metody estymacji gęstości. W dalszej kolejności prezentuję ilościową ocenę odporności zaproponowanej metody analizy skupisk na szum addytywny i szum tła. Przedstawiam także wyniki analogicznej oceny ilościowej dla trzech zaproponowanych w literaturze, znaczących algorytmów analizy skupisk w danych obarczonych szumem. Kolejne wyniki eksperymentalne wykazują zbieżność zaproponowanej metody wizualizacji danych i prezentują ilościową ocenę jakości odwzorowań konstruowanych z wykorzystaniem zaproponowanej heurystycznej techniki przyspieszania procesu wizualizacji.

Kolejna część rozprawy przedstawia wyniki, które uzyskałem stosując zaproponowane algorytmy w rzeczywistych zagadnieniach związanych z analizą skupisk w danych biomedycznych. Prezentuję wyniki dla czterech rożnych typów danych. Pierwsze rezultaty przedstawiają identyfikację skupisk w sekwencjach niekodujących cząsteczek RNA. Następnie przedstawiam analizę skupisk w zbiorze trójwymiarowych struktur łańcuchów białkowych występujących w komórkach człowieka. W dalszej kolejności prezentuję zastosowanie zaproponowanej metody estymacji gęstości w zagadnieniu wyboru z dużej ilości danych mikromacierzowych niewielkiego zestawu profili ekspresji genów, który precyzyjnie opisuje zjawisko biologiczne będące przedmiotem eksperymentu mikromacierzowego. Na zakończenie tej części pracy przedstawiam analizę skupisk w zbiorze wielowymiarowych wektorów cech opisujących fragmenty zdjęć monograficznych, w których podejrzewa się występowanie tzw. *mikrozwapnień*. W każdym omawianym zastosowaniu prezentowane wyniki odnoszone są do dostępnej wiedzy dziedzinowej, pozyskanej z publicznych repozytoriów danych biomedycznych.

Pracę zamyka rozdział podsumowujący najważniejsze wnioski płynące z uzyskanych rezultatów i odnoszący je do postawionej tezy oraz celów pracy. W rozdziale tym omawiam także szczególnie obiecujące kierunki dalszych badań nad algorytmami analizy skupisk i wizualizacji zbiorów danych obarczonych szumem.

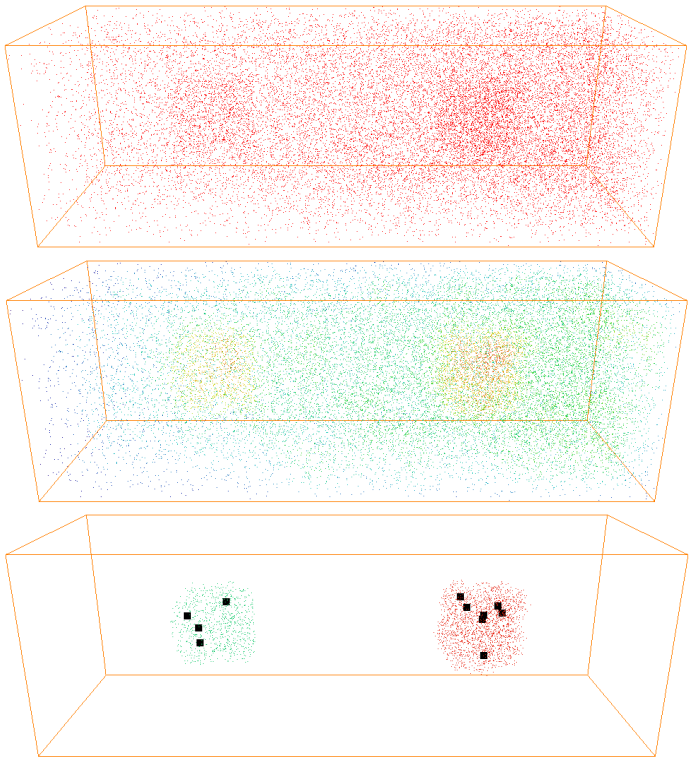
Najważniejsze wyniki

Prezentację oryginalnych rezultatów rozprawy rozpoczyna rozdział 5. W pierwszej części rozdziału proponuję nową metodę estymacji rozkładów gęstości w zbiorach obserwacji. W algorytmie tym obserwacje są organizowane są w graf k–najbliższych sąsiadów. Następnie, wykonywana jest iteracyjna procedura szacowania poprawek do poziomów gęstości, która buduje rozkład gęstości obserwacji w zbiorze danych. Procedura estymacji gęstości stanowi podstawę algorytmu analizy skupisk, który zaproponowałem w drugiej części rozdziału 5. Algorytm ten rozpoczyna analizę skupisk usuwając ze zbioru danych obserwacje, dla których oszacowano wartości gęstości poniżej zadanej przez użytkownika wartości progowej. Pośród pozostałych obserwacji algorytm identyfikuje obserwacje reprezentatywne. Wokół obserwacji reprezentatywnych algorytm konstruuje skupiska homocentryczne (ang. *homocentric clusters*). Skupiska te mogą być następnie łączone w skupiska wielocentryczne (ang. *multicentric clusters*), jeśli w zbiorze danych poszukiwana jest „gruboziarnista" struktura skupisk. Końcowa część rozdziału 5 poświęcona jest szczegółom implementacji zaproponowanych algorytmów. Prezentuję, miedzy innymi, implementacje równoległe zaproponowanych algorytmów, przeznaczone dla komputerów wieloprocesorowych z pamięcią współdzieloną.

Rozdział 6 prezentuje rezultaty rozprawy w zakresie wizualizacji zbiorów danych. Proponuję w nim rozwinięcie opisanej w literaturze cząsteczkowej metody skalowania wielowymiarowego. Metoda wyjściowa konstruuje odwzorowanie wizualizowanego zbioru danych poprzez prowadzenie klasycznej dynamiki cząstek (reprezentujących obserwacje w zbiorze danych) z funkcją potencjału dla oddziaływań konserwatywnych zadaną przez funkcję naprężenia wizualizacji oraz z uwzględnieniem uproszczonych sił tarcia, które rozpraszają energię kinetyczną. Proponuję rozszerzenie polegające na zastąpieniu klasycznej dynamiki cząstek Dyssypatywną Dynamiką Cząstek (ang. *Dissipative Particle Dynamics – DPD*), co pozwala na jawną kontrolę temperatury układu w trakcie procesu poszukiwania minimum. Dzięki temu możliwe staje się symulowane wyżarzenie układu. W rozdziale 6 prezentuję także heurystyczną metodę przyspieszenia procesu symulowanego wyżarzania. Polega ona na usunięciu części członów z funkcji naprężenia wizualizacji. Dzięki temu maleje ilość oddziaływań w symulowanym układzie, a tym samym maleje ilość obliczeń przypadająca na krok symulacji. W końcowej części rozdziału 6 prezentuję implementację równoległą zaproponowanej metody wizualizacji, przeznaczoną dla komputerów wieloprocesorowych z pamięcią dzieloną.

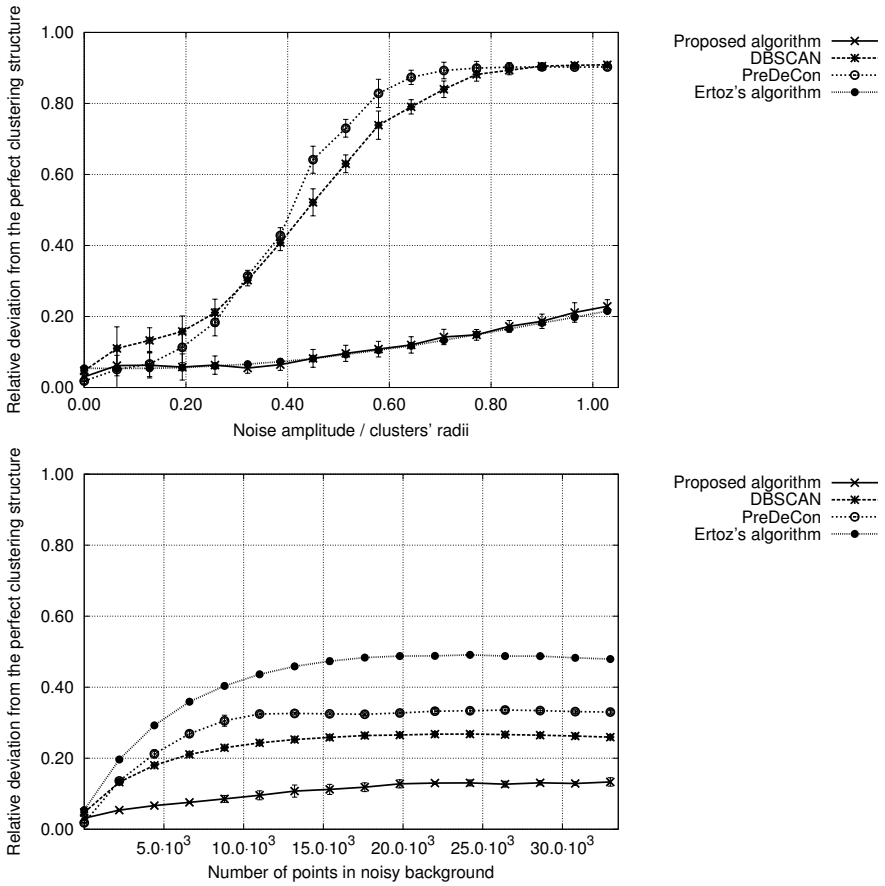
W rozdziale 7 przedstawiam wyniki eksperymentalnej oceny skuteczności zaproponowanych algorytmów. Rozdział otwiera opis zaczerpniętej z literatury metody wyznaczania odległości pomiędzy dwoma propozycjami struktury skupisk dla tego samego zbioru danych. Omawiam także zaczerpniętą z literatury metodę dopasowania różnych wizualizacji tego samego zbioru danych oraz jej zastosowanie do wyznaczania odległości pomiędzy wizualizacjami. Następnie prezentuję testowy, syntetyczny zbiór danych składający się z 11 częściowo przylegających do siebie skupisk z 10–cio wymiarowej przestrzeni Euklidesowej. Omawiam także technikę wprowadzania do tych danych szumu addytywnego oraz szumu tła. Omówione tu miary odległości i testowy zbiór danych są wykorzystane we wszystkich ilościowych testach prezentowanych w dalszej części rozdziału 7.

Ocenę skuteczności zaproponowanego algorytmu analizy skupisk rozpoczynam od zaprezentowania kilku przykładowych wyników dla dwu- i trójwymiarowych zbiorów danych. Jeden z prezentowanych rezultatów przedstawiono na rysunku 1. W dalszej części roz-



Rysunek 1: Rezultaty analizy skupisk w syntetycznym, trójwymiarowym zbiorze danych, składającym się z dwóch sześciennych skupisk „zanurzonych" w intensywnym szumie tła. Przedstawiono oryginalny zbiór danych, rozkład gęstości obserwacji oszacowany za pomocą zaproponowanego w rozprawie estymatora gęstości oraz skupiska zidentyfikowane za pomocą zaproponowanego w rozprawie algorytmu analizy skupisk.

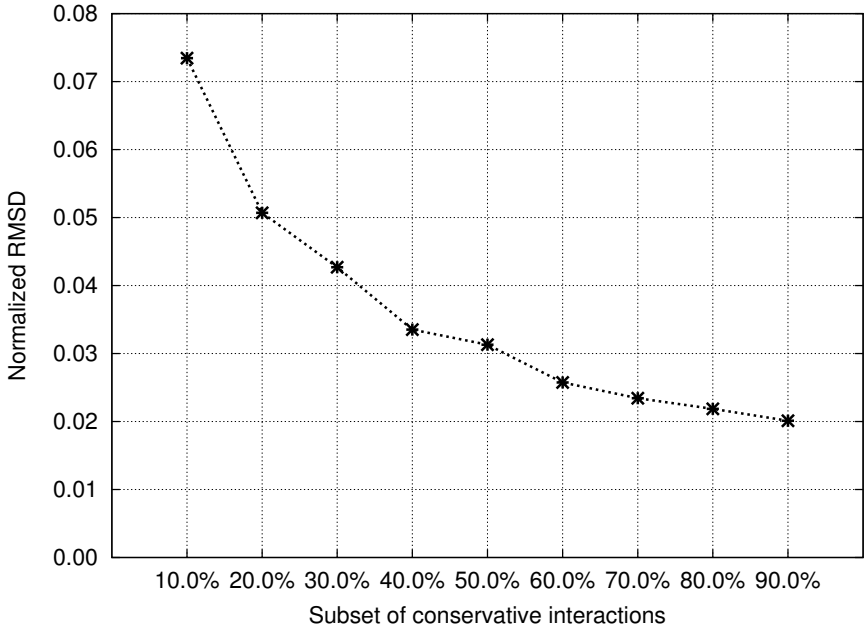
działu 7 prezentuję wyniki eksperymentalne wskazujące, iż procedura estymacji gęstości jest zbieżna. Rezultat ten jest istotny ze względu na iteracyjny charakter tej procedury. Kolejne wyniki prezentują ilościowo odporność zaproponowanego algorytmu analizy skupisk na szum addytywny i szum tła. Wyniki te uzyskałem wprowadzając do testowego zbioru danych szum o rosnącej intensywności i odnosząc, za pomocą wspomnianej powyżej miary odległości, struktury skupisk konstruowane przez zaproponowany algorytm do znanej, prawidłowej struktury skupisk. Analogiczną ocenę ilościową przeprowadziłem dla trzech zaczerpniętych z literatury, znaczących algorytmów analizy skupisk w danych obarczonych szumem. Uzyskane rezultaty przedstawiają wykresy na rysunku 2. Rygorystyczna ocena ilościowa skuteczności porównywanych metod ujawnia, iż algorytmy zaczerpnięte z literatury są w istocie wrażliwe na rosnący poziom szumu. W przeciwieństwie do nich, algorytm który zaproponowałem w rozprawie wykazuje wysoką jakość konstruowanych struktur skupisk dla szerokiego zakresu wartości intensywności szumu.



Rysunek 2: Wyniki ilościowej oceny odporności porównywanych algorytmów analizy skupisk na szum addytywny i szum tła. Spośród porównywanych metod, jedynie algorytm zaproponowany w omawianej rozprawie wykazuje odporność zarówno na szum tła jak i szum addytywny w szerokim zakresie ich intensywności.

Ostatnia część rozdziału 7 poświęcona jest testom zaproponowanego rozwinięcia cząsteczkowej metody skalowania wielowymiarowego. Pokazuję, iż zastosowanie DPD istotnie umożliwia kontrolę temperatury w układzie. Prezentuję symulowane wyżarzanie dla zagadnienia wizualizacji omówionego powyżej 10–cio wymiarowego zbioru testowego, demonstrując zbieżność zaproponowanego algorytmu. Następnie, prezentuję ilościową ocenę jakości wizualizacji konstruowanych z wykorzystaniem heurystycznej metody przyspieszania procesu wyżarzania. Oceny tej dokonałem poprzez porównanie wizualizacji skonstruowanych z wykorzystaniem funkcji naprężenia z której usunięto część członów z wizualizacją otrzymaną przy zastosowaniu pełnej funkcji naprężenia. Uzyskane rezultaty przedstawia wykres na rysunku 3. Usunięcie nawet 90% członów z funkcji naprężenia skutkuje wystąpieniem względnego średniego błędu kwadratowego wizualizacji mniejszego niż 8%.

Rozdział 7 prezentuje także ocenę wydajności zaproponowanych implementacji równoległych omawianych algorytmów. W przypadku algorytmu analizy skupisk uzyskano ok. 54–krotne przyspieszenie dla 64 procesorów. Równoległa implementacja algorytmu wizualizacji umożliwia uzyskanie około 53–krotnego przyspieszenia dla 64 procesorów. W obu



Rysunek 3: Względny średni błąd kwadratowy wizualizacji zbioru danych przy zastosowaniu funkcji naprężenia, z której usunięto część członów, w stosunku do wizualizacji otrzymanej przy wykorzystaniu pełnej funkcji naprężenia.

przypadkach wydajność zrównoleglenia przy 64 procesorach wynosi więc ok. 80%.

W rozdziale 8 rozprawy prezentuję zastosowania zaproponowanych algorytmów w czterech różnych problemach badawczych, związanych z analizą skupisk w danych biomedycznych. Rozdział rozpoczynam wynikami analizy skupisk w zbiorze złożonym z około 32,000 sekwencji niekodujących cząsteczek RNA. Sekwencje te zaczerpnięto z bazy danych Rfam klasyfikującej niekodujące RNA do kilkuset różnych rodzin. Większość rodzin zawiera niewielką ilość cząsteczek RNA i w omawianej analizie stanowi szum tła. Pozostałe, większe rodziny powinny natomiast tworzyć skupiska. Analizę skupisk wykonałem na podstawie macierzy wartości miary niepodobieństwa sekwencji wyznaczonej z punktacji dopasowań par sekwencji (ang. *pairwise sequence alignment scores*). Wartości parametrów algorytmu analizy skupisk dobrałem kierując się trójwymiarową reprezentacją zbioru sekwencji skonstruowaną za pomocą zaproponowanej metody wizualizacji danych. Algorytm analizy skupisk skonstruował 42 skupiska i dwie niewielkie grupy sekwencji ncRNA o wartościach gęstości bliskich wartości progowej dla obserwacji w szumie tła. Dla każdego z 42 skupisk wyznaczyłem *czystość* skupisk rozumianą jako odsetek sekwencji w skupisku przynależnych do najliczniej reprezentowanej rodziny. Spośród skonstruowanych skupisk 37 wykazuje czystość przekraczającą 99%, trzy wykazują czystość przekraczającą 90%, zaś kolejne dwa wykazują czystość bliską, odpowiednio, 88% i 74%. W podobny sposób przeprowadziłem analizę skupisk w zbiorze złożonym z około 16,100 trzeciorzędowych struktur łańcuchów białkowych syntetyzowanych w komórkach organizmu ludzkiego. Macierz niepodobieństwa struktur łańcuchów wyznaczyłem za pomocą oprogramowania *DaliLite*, służącego do dopasowywania struktur białek. Algorytm analizy skupisk skonstruował 58 skupisk. Skupiska

te zostały porównane z klasyfikacją zwojów białkowych udostępnianą przez bazę SCOP. Zgodnie z tą klasyfikacją, 54 skupiska wykazują czystość na poziomie 100% zaś kolejne dwa skupiska wykazują czystość na poziomie, odpowiednio, 84% i około 81.5%. Ostatnim dwóm skupiskom, zawierającym w sumie jedynie 51 łańcuchów białkowych, nie odpowiadał żaden zwój białkowy z klasyfikacji SCOP.

Kolejne rezultaty w rozdziale 8 dotyczą zagadnienia selekcji z danych mikromacierzowych niewielkiego zestawu profili ekspresji genów, precyzyjnie opisującego procesy biologiczne badane w eksperymencie mikromacierzowym. Zaproponowałem, by selekcję profili ekspresji genów prowadzić na podstawie kryterium gęstości oszacowanej algorytmem omawianym w rozdziale 5. Podejście to zostało porównane z klasyczną metodą selekcji podzbioru profili ekspresji, wykorzystującą kryterium odchylenia standardowego poziomów ekspresji. Porównanie przeprowadzono szacując skuteczność klasyfikacji pięciu różnych typów raka płuc oraz zdrowej tkanki płucnej na podstawie wyselekcjonowanych profili ekspresji. Rezultaty wskazują, iż zaproponowana metoda selekcji profili ekspresji pozwala uzyskać wyższą dokładność klasyfikacji.

Ostatni problem poruszany w rozdziale 8 dotyczy detekcji skupisk w wektorach cech opisujących fragmenty zdjęć monograficznych podejrzewane o obecność mikrozwapnień. Algorytmy zaproponowane w prezentowanej rozprawie pozwoliły wykryć 5 skupisk w zbiorze składającym się z około 93,000 27-mio wymiarowych wektorów cech. Uzyskane skupiska poddano wizualnej ocenie pod kątem rodzaju widocznych na nich struktur radiologicznych. Następnie, dla każdego skupiska wyznaczono najliczniej reprezentowany rodzaj struktury radiologicznej i oszacowano jego czystość. Jedno ze skupisk wykazuje czystość na poziomie ok. 99%, kolejne dwa na poziomie około 94%, zaś dwa ostatnie na poziomie odpowiednio 91% i 84.5%.

Podsumowanie i wnioski

Rezultaty moich prac badawczych wskazują, iż możliwa jest estymacja rozkładu gęstości obserwacji w zbiorze danych na podstawie pozycji obserwacji listach sąsiedztwa grafu k–najbliższych sąsiadów. Tak sformułowana metoda estymacji gęstości może być stosowana dla dowolnego rodzaju obserwacji, które można porównywać miarą niepodobieństwa. Obserwacje te nie muszą więc mieć charakteru danych wektorowych. W rozprawie wykazałem, iż opracowany estymator gęstości może być z powodzeniem zastosowany do konstrukcji skutecznego algorytmu analizy skupisk, korzystającego z kryterium lokalnej gęstości obserwacji. W szczególności, zaproponowałem algorytm analizy skupisk, który, w odróżnieniu od kilku znaczących metod tej klasy opublikowanych w literaturze, jest odporny na szum addytywny i szum tła w szerokim zakresie ich intensywności.

Zaproponowany w rozprawie algorytm wizualizacji zbiorów danych łączy cechy cząsteczkowej metody skalowania wielowymiarowego z metodami symulowanego wyżarzania. Algorytm ten odwzorowuje zbiory danych w przestrzeni Euklidesowej poprzesz minimalizację funkcji naprężenia wizualizacji. Ponieważ funkcja naprężenia zadana jest jedynie przez konfigurację cząsteczek w układzie oraz macierz wartości miary niepodobieństwa dla par obserwacji, zaproponowany algorytm może być stosowany dla dowolnych rodzajów obserwacji, dla których dostępna jest miara niepodobieństwa.

Rezultaty, które uzyskałem w rozdziale poświęconym analizie danych biomedycznych potwierdzają skuteczność metod, które zaproponowałem. W rozdziale tym pokazałem, iż skupiska konstruowane za pomocą zaproponowanych algorytmów wykazują wysoką zgodność z istniejącymi klasyfikacjami obiektów biologicznych. Tym samym, skupiska te oddają naturę badanych obiektów i procesów biologicznych.

Część z uzyskanych rezultatów wskazuje na interesujące kierunki dalszych badań. Szczególnie zachęcające wydają się być wyniki wizualizacji zbiorów danych z wykorzystaniem funkcji naprężenia, z której usunięto część członów (rysunek 5). Usunięcie nawet 90% członów z funkcji naprężenia skutkuje jedynie niewielkim pogorszeniem jakości wizualizacji. Wydaje się, iż adaptacyjne usuwanie członów z funkcji naprężenia, uwzględniające np. zależności geometryczne, może prowadzić do jeszcze lepszych rezultatów. Rozwojowy charakter ma także zastosowanie metody estymacji gęstości do problemu selekcji podzbioru profili ekspresji z danych mikromacierzowych. Podejście to może okazać się przydatne także w innych zagadnieniach, w których pojawia się problem nienadzorowanej selekcji cech. Obiecujące wydają się być także badania zmierzające do zmniejszenia złożoności obliczeniowej zaproponowanych metod. W zaproponowanym algorytmie analizy skupisk najbardziej kosztownym obliczeniowo jest wyznaczanie najkrótszych ścieżek w grafie najbliższych sąsiadów. W literaturze dostępne są metody szacowania przybliżonych długości najkrótszych ścieżek w grafie. Mogą one zmniejszyć złożoność obliczeniową zaproponowanego algorytmu. Ich stosowanie wymaga jednak uprzedniej weryfikacji jakości konstruowanych w ten sposób skupisk.

Wybrane publikacje doktoranta

1. K. Boryczko, M. Kurdziel *Approximate Clustering of Noisy Biomedical Data*, Lecture Notes in Computer Science, vol. 5101, pp. 630–640, Springer–Verlag, 2008

2. M. Kurdziel, K. Boryczko, D.A. Yuen *Detecting Clusters of Microcalcifications in High–Resolution Mammograms Using Support Vector Machines*, Bio–Algorithms and Med–Systems, vol. 3(6), pp. 11–22, CM UJ, 2007

3. T. Arodź, M. Kurdziel, T.J. Popiela, E.O.D. Sevre, D.A. Yuen *Detection of Clustered Microcalcifications in Small Field Digital Mammography*, Computer Methods and Programs in Biomedicine, vol. 81(1), pp. 56–65, Elsevier Science, 2006

4. T. Arodz, K. Boryczko, W. Dzwinel, M. Kurdziel, D.A. Yuen *Visual Exploration of Multidimensional Feature Space of Biological Data*, in: "Proc. IEEE Visualization 2005", pp. 90, IEEE Computer Society, 2005, abstract

5. T. Arodź, M. Kurdziel, E.O.D. Sevre, D.A. Yuen *Pattern Recognition Techniques for Automatic Detection of Suspicious–looking Anomalies in Mammograms*, Computer Methods and Programs in Biomedicine, vol. 79(2), pp. 135–149, Elsevier Science, 2005

6. K. Boryczko, M. Kurdziel *Recognition of Subtle Microcalcifications in High–Resolution Mammograms*, in: "Proc. IV Int'l Conf. on Computer Recognition Systems, CORES

2005", Advances in Soft Computing, vol. 30, pp. 485–492, Springer–Verlag, 2005

7. K. Boryczko, M. Kurdziel *Parallel Clustering of Large–Scale Noisy Multidimensional Datasets*, in: "Proc. Cracow Grid Workshop '04", ACC Cyfronet AGH, Kraków, 2005

8. W. Dzwinel, K. Boryczko, T. Arodź, M. Kurdziel *Komputerowe Metody Detekcji Nowotworów Piersi w Zdjęciach Mammograficznych*, Bio–Algorithms and Med–Systems, vol. 1(1/2), pp. 287–290, CM UJ, 2005

9. T.J. Popiela, A. Urbanik, T. Arodź, M. Kurdziel *Computer–aid system for the detection of clustered microcalcifications in digital mammography*, Polish Journal of Radiology, vol. 69(S1), pp. 67–68, 2004, abstract