Politechnika Śląska w Gliwicach
Wydział Automatyki, Elektroniki i Informatyki

Autoreferat rozprawy doktorskiej

Bi-klasteryzacja – algorytmy i ich zastosowania

Paweł Foszner

Praca przygotowana pod kierunkiem
prof. dr hab. inż. Andrzej Polański

Gliwice, 2014

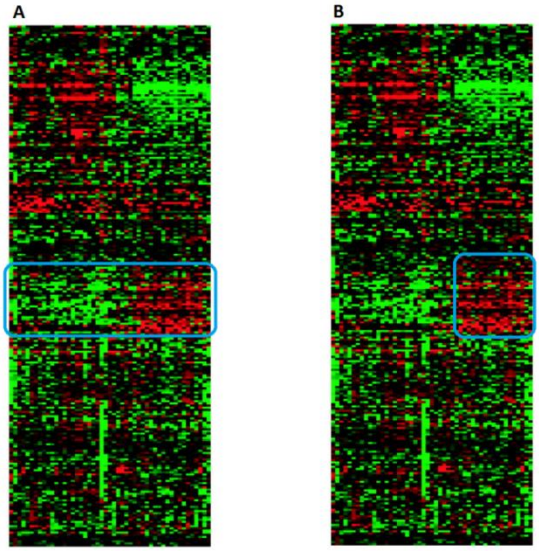
Spis treści

1. Wprowadzenie 3
2. Cele i tezy pracy 6
3. Środowisko obliczeniowe 7
4. Wybrane wyniki obliczeniowe 8
5. Podsumowanie 12
6. Publikacje autora 13

Bibliografia 15

1. Wprowadzenie

W obecnych czasach stale obserwujemy szybki rozwój w dziedzinie telemetrii, analiz biomedycznych, analiz tekstowych oraz szeroko pojętej eksploracji danych. Rezultatem tych badań są zwykle duże oraz złożone zestawy danych. Klasyczne oraz dobrze udokumentowane techniki pozwalają na wydobycie jedynie części istotnych informacji. Rysunek 1 przedstawia macierz ekspresji genów pochodzącą z serii eksperymentów mikro-macierzowych. Pojedyncza komórka przedstawia ekspresję danego genu pod ustalonymi warunkami eksperymentu.

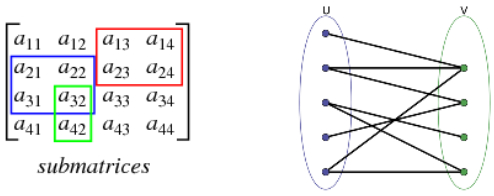


Rysunek 1. Porównanie klasycznej klasteryzacji z bi-klasteryzacją.

Klasyczne podejście do klasteryzacji danych, przedstawione na obrazie (A) pozwoli na znalezienie grupy genów podobnych do siebie z punktu widzenia wszystkich warunków umieszczonych w danych. Obraz (B) na przedstawia dokładnie tą samą macierz danych poddaną analizie bi-klasteryzacji. Oprócz podobnej grupy genów eksperyment wskazał także grupę warunków pod którymi ta grupa wykazuje podobieństwo.

Przytoczony przykład obrazuje różnice między klasycznym i uniwersalnym podejściem a jej bardziej złożonym i wyspecjalizowanym odpowiednikiem. Bi-klasteryzacja jest to wyszukiwanie w dwuwymiarowej przestrzeni danych podzbioru atrybutów z jednego wymiaru który wykazuje pewne podobieństwo tylko wśród

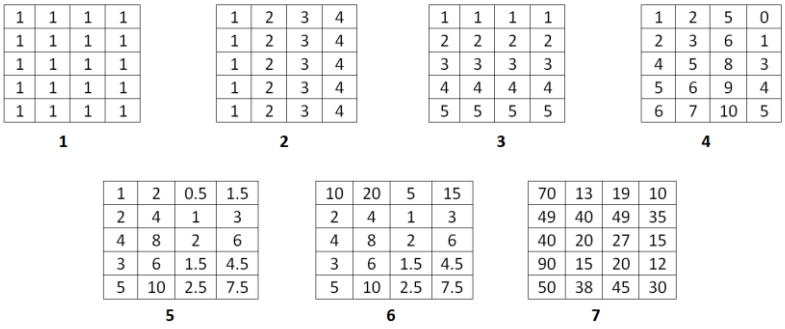
podzbioru atrybutów z wymiaru drugiego. W bardzo prostych słowach można opisać bi-klasteryzację jako wyszukiwanie podmacierzy w macierzy danych lub klik w grafie dwudzielnym (Rysunek 2).



Rysunek 2. Prosta wizualizacja bi-klasteryzacji.

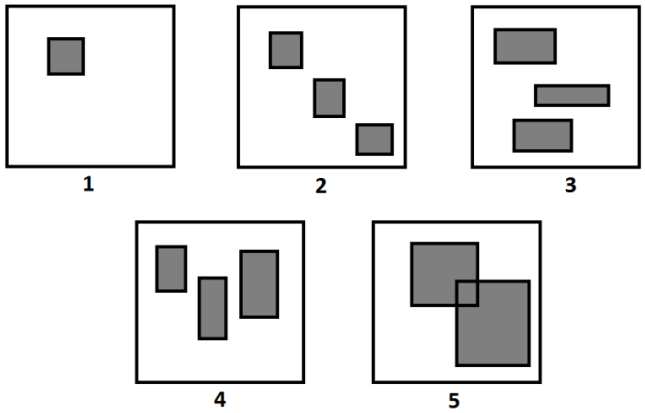
Bi-klasteryzacja, jak już przytoczono powyżej, jest techniką eksploracji danych za pomocą której można dokonać współbieżnej klasteryzacji kolumn oraz wierszy macierzy danych. Technika ta należy do klasy problemów NP-trudnych oraz została po raz pierwszy poruszona w literaturze w 1963 roku przez Morgana i Sonquista [1], następnie przez Hartigana [2] (1972) oraz Mirkina w 1992 [3] . W kontekście analizy bio-informatycznej pierwsze udokumentowane użycie tej techniki można znaleźć w pracach Cheng'a i Church'a [4]. Autorzy jako pierwsi użyli technik bi-klasteryzacji do analizy danych mikro-macierzowych. Na przestrzeni ostatnich 50 lat powstało wiele różnych podejść do omawianego zagadnienia. Metody różnią się między sobą zarówno w podejściu do modelowania danych wejściowych (graf dwudzielny [5], macierz dyskretna [6], drzewa [7]), a także sposobie uzyskiwania ostatecznych wyników (wyczerpujące wyszukiwanie [5], rozkład macierzy [4], przeszukiwanie grafu [2]).

Wyróżnia się wiele rodzajów bi-klastrów ze względu na ich strukturę oraz położenie. Rysunek 3 przedstawia klasyfikację bi-klastrów ze względu na dane. Wyróżniamy klastry o stałych wartościach, o wartościach przesuniętych lub przeskalowanych. Wiele metod istniejących w literaturze specjalizuje się najczęściej w jednym rodzaju struktury przedstawionym na poniższym rysunku.



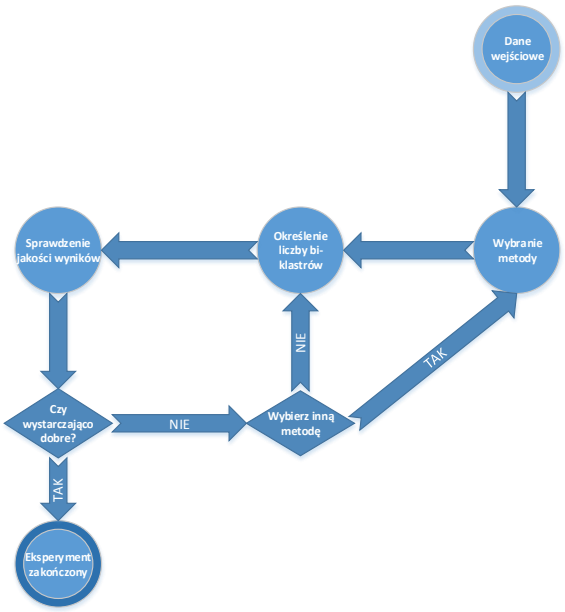
Rysunek 3. Prosta wizualizacja prezentująca różne rodzaje struktury bi-klastrów.

Innym bardzo istotnym podziałem bi-klastrów jest podział ze względu na położenie klastra w macierzy danych. Wyróżniamy dane gdzie w żadnym wymiarze poszczególne klastry nie nachodzą na siebie, takie gdzie nachodzą w ramach jednego wymiaru lub obu jednocześnie.



Rysunek 4. Klasyfikacja bi-klastrów ze względu na położenie w macierzy danych.

Zdecydowana większość algorytmów w literaturze bierze pod uwagę wspomniane wyżej cechy oraz dostosowuje pod nie swoje działanie. W rezultacie przeprowadzając eksperymenty bi-klasteryzacji należy być w pełni świadomym posiadanych danych, ich struktury oraz spodziewanej liczby i wielkości bi-klastrów. Jako że niemal nigdy nie można z całkowitą pewnością określić tych cech, typowy eksperyment bi-klasteryzacji wygląda jak ten opisany na rysunku 5. Jest to proces iteracyjny, w którym ręcznie dobiera się metody oraz parametry tak aby uzyskać jak najlepszą jakość wyników.



Rysunek 5. Diagram przepływu danych w typowym eksperymencie bi-klasteryzacji.

2. Cele i tezy pracy

2.1. Cele

Główne cele stworzonej pracy to:

- Stworzenie publicznie dostępnego oraz darmowego systemu zawierającego wszystkie najważniejsze algorytmy bi-klasteryzacji.
- Przetestowanie i porównanie tych algorytmów na danych rzeczywistych oraz syntetycznych.
- Wprowadzenie ulepszeń do procesu bi-klasteryzacji polegających na wyeliminowaniu wpływu struktury danych na wyniki końcowe. Zaproponowany algorytm polega na umiejętnym łączeniu wyników metod wyspecjalizowanych w znajdywaniu bi-klastrów o różnej strukturze.

Głównym celem prowadzonych badań było za pomocą nowego algorytmu oraz systemu agregującego różne metody tak uprościć analizę bi-klasteryzacji aby uzyskanie satysfakcjonujących wyników ograniczało się do podania danych na wejściu (Rysunek 6). Kluczowym elementem zaproponowanej metody jest wykonanie możliwie jak największej liczby eksperymentów bi-klasteryzacji wyspecjalizowanych w możliwej jak największej liczbie danych. Następnie wszyst-

kie wyniki łączone są w jeden. W tym celu zaproponowano miary podobieństwa bi-klastrów oraz zaproponowano modyfikacje algorytmu węgierskiego w celu parowania ich.

Rysunek 6. Diagram uproszczonej analizy bi-klasteryzacji.

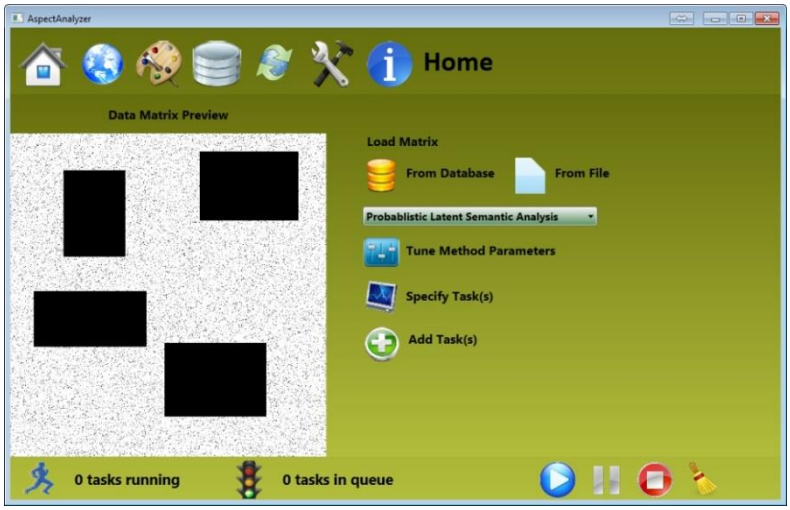
2.2. Tezy

Główne tezy pracy to:

- Opracowana metodologia porównywania wyników bi-klasteryzacji oparta na uogólnionym algorytmie Munkres'a pozwala na dokonywanie porównań zarówno w przypadku bi-klasteryzacji ze znanym wzorcem jak tez w przypadku, gdy prawdziwa struktura bi-klastrów jest nieznana.
- Opracowany nowy algorytm łączenia wyników bi-klasteryzacji (meta-algorytm) pozwala na uzyskanie poprawy jakość bi-klasteryzacji.

3. Środowisko obliczeniowe

Na potrzeby analiz oraz eksperymentów opisanych w rozprawie stworzono graficzny i intuicyjny system do obsługi nowych algorytmów. Rysunek 7 przedstawia główne okno narzędzia.



Rysunek 7. Główne okno systemu do automatycznej klasteryzacji.

System jest oparty o platformę .NET Framework oraz został napisany w języku C#. Graficzny interfejs bazuje na Windows Presentation Foundation i można go uruchomić w dowolnym środowisku Windows które obsługuje .NET Framework w wersji 4.5 lub wyższej. Obliczenia matematyczne oparte są na otwartej bibliotece ILNumerisc [8].

Oprogramowanie ma możliwość uruchamiania i zarządzania obliczeniami rozproszonymi, zarówno w ramach jednego procesora obsługującego wiele wątków, jak i w ramach wielu węzłów rozproszonych w sieci.

Zarówno cały system jak i komponenty od których jest zależny (poza systemem operacyjnym) są darmowym oprogramowaniem, oraz został opublikowany na stronie http://aspectanalyzer.foszner.pl.

4. Wybrane wyniki obliczeniowe

4.1. Dane syntetyczne.

Na potrzeby analiz danych syntetycznych zostało stworzonych 54 różne macierze reprezentujące różne typy macierzy danych. Ze względu na strukturę danych wyróżniono następujące typy:

- Dane o stałych wartościach (0),
- Dane o stałych wartościach (> 0),
- Dane przesunięte,
- Dane przeskalowane,
- Dane przesunięte oraz przeskalowane,
- Dane warstwowe (plaid data).

Natomiast ze względu na położenie oraz ilość bi-klastrów wyróżniamy następujące typy:

- Dane zawierające pojedynczy bi-klaster,
- Dane z bi-klastrami nie nachodzącymi na siebie w żadnym wymiarze,
- Dane z nachodzącymi na siebie kolumnami (do 25%),

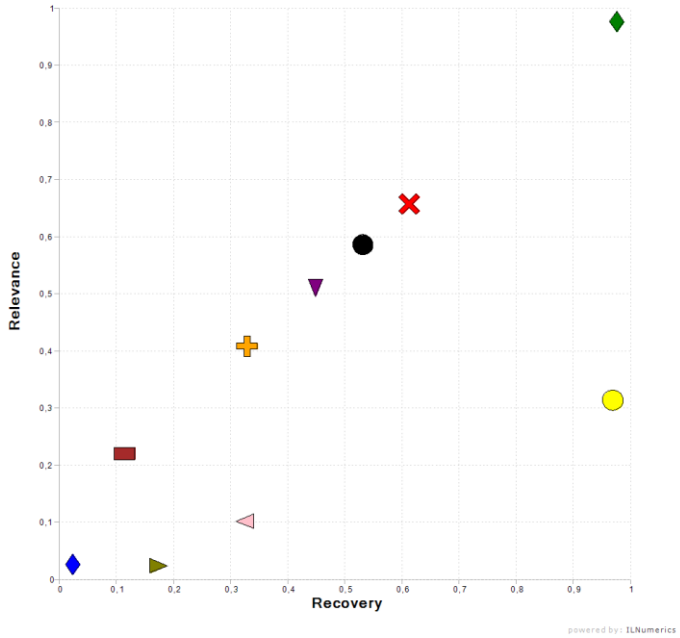
- Dane z nachodzącymi na siebie kolumnami (do 50%),
- Dane z nachodzącymi na siebie kolumnami (do 75%),
- Dane z nachodzącymi na siebie wierszami (do 25%),
- Dane z nachodzącymi na siebie wierszami (do 50%),
- Dane z nachodzącymi na siebie wierszami (do 75%),
- Dane z klastrami nachodzącymi na siebie oboma wymiarami (do 100%).

Pojedynczy wynik dla danych syntetycznych badany był pod dwoma względami: (1) Czy znaleziono wszystkie oczekiwane bi-klastry (recovery) oraz (2) Czy wszystkie znalezione bi-klastry były oczekiwane (relevance). Obie te miary przyjmują wartości z przedziału <0, 1> oraz w idealnym przypadku oczekujemy wartości 1 dla każdej z nich. Jako miarę jakości dla algorytmu przyjęto wartość średnią z (1) oraz (2).

Każda macierz została poddana analizie za pomocą jednego z dziewięciu popularnych algorytmów z literatury. Poniżej ich lista wraz z symbolem jaki reprezentują na wykresie:

- BBC ●
- Cheng-Church ✖
- BiMax ♦
- CPB ●
- FABIA ✚
- XMotifs ▬
- Plaid ▼
- ISA ◀
- Qubic ▶

Po wykonaniu wszystkich eksperymentów, ich wyniki zostały przefiltrowane oraz połączone w jeden. Wynik takiej analizy jest przestawiony na wykresie symbolem ♦. Spośród 54 niezależnych porównań, niemal wszystkie wykazały, że podejście zaproponowane w rozprawie znacząco poprawia automatyczną analizę.



Rysunek 8. Przykładowe wyniki dla macierzy z jednym bi-klasterem którego wartości są przeskalowane i przesunięte.

Poniżej w tabeli 1 przedstawiono rozszerzony opis dla rysunku 8. W tabeli wyszczególniony został algorytm który uzyskał najlepszy (zielony) oraz najgorszy wynik (czerwony).

Tabela 1. Przykładowe wyniki dla macierzy z jednym bi-klasterem którego wartości są przeskalowane i przesunięte.

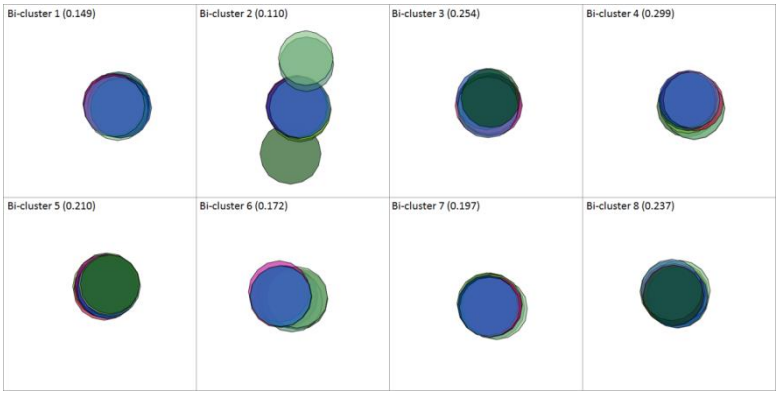
Method name	Chart symbol	Recovery	Relevance	Score	Average Num. of bi-clusters
BBC	●	0,594	0,594	0,594	1
Cheng-Church	✖	0,669	0,669	0,669	1
BiMax	◆	0,131	0,013	0,072	10
CPB	●	0,994	0,312	0,653	1
FABIA	✚	0,41	0,41	0,41	4,12
XMotifs	▬	0,214	0,214	0,214	1
Plaid	▼	0,519	0,519	0,519	1
ISA	◀	0,409	0,091	0,25	4,57
Qubic	▶	0,266	0,011	0,138	25,49
Consensus	◆	1	1	1	1

4.2. Dane rzeczywiste.

Do porównania wzięto 2 rodzaje danych:

- Dane bio-informatyczne – macierz ekspresji genów pochodząca z eksperymentów mikro-macierzowych przeprowadzonych na 6 różnych grupach pacjentów chorych ba białaczkę. Oczekiwane rezultaty to 6 bi-klastrów złożonych z grupy genów oraz grupy warunków reprezentujących jedną z grup pacjentów każdy.
- Dane tekstowe – macierz wystąpień słów w kontekście genów, powstała na bazie 1071 artykułów opisujących jeden z ośmiu terminów ontologii genowych. Oczekiwane rezultaty to 8 bi-klastrów złożonych z grupy genów oraz grupy słów reprezentujących jeden z terminów ontologii genowych każdy.

Uzyskane wyniki były analizowane pod kątem jakości otrzymanych wyników. Miarą tej jakości był współczynnik ACV (Average Corelation Value). Po wykonaniu serii eksperymentów dla wielu różnych metod, pierwszym istotnym elementem analizy jest odfiltrowanie elementów odstających. W tym celu bi-klastry pochodzące z różnych wyników są parowane oraz nakładane na siebie. Odrzucane są wyniki odstające od większości (Rysunek 9).



Rysunek 9. Analiza powtarzalności wyników w ramach pojedynczeń metody.

Po odfiltrowaniu bi-klastrów odstających znacząco od reszty następuje ich grupowanie. Utworzone grupy składają się z bi-klastrów będących swoimi odpowiednikami między różnymi eksperymentami. Następnie w ramach każdej grupy bi-klastry scalane są tak aby utworzyły jeden bi-klaster. Parametry takie jak ilość bi-

klastrów, ich wielkość lub pokrycie mogą być zarówno parametrami podanymi przez użytkownika, jak również mogą zostać wyszukane przez algorytm.

Tabela 2 przedstawia przykładowe wyniki dla 4 różnych metod. W ramach każdej z nich wykonano serię eksperymentów oraz podano średnią oraz najlepszą wartość miary jakości bi-klastra. Następnie każdy z tych wyników został podany na wejście meta-algorytmu. Ostatnim etapem była analogiczna analiza wszystkich wyników. Uzyskane wartości pozwalają jednoznacznie stwierdzić że podejście oparte na meta-algorytmie znacząco poprawia jakość prowadzonych analiz. Wskazują na to poprawione wartości miary jakości w każdym przypadku.

Tabela 2. Przykładowe wyniki dla danych pochodzących z analiz tekstowych.

Method	Type	Average AVC	Best AVC
PLSA	Normal	0.118	0.138
	Consensus	0.304	
K-L	Normal	0.129	0.147
	Consensus	0.297	
LSE	Normal	0.211	0.245
	Consensus	0.274	
nsK-L	Normal	0.140	0.154
	Consensus	0.253	
All results	Normal	0.186	0,233
	Consensus	0.345	

5. Podsumowanie

Celem pracy było opracowanie uniwersalnego podejścia do analiz bi-klasteryzacji oraz uodpornienie metody na strukturę posiadanych danych. W tym celu stworzono zbiór danych syntetycznych który pokrywał niemal wszystkie istotne warianty danych. Uzyskane na ich podstawie wyniki wykazały że podejście zaproponowane w rozprawie jest wyraźnie lepsze od dostępnych metod lub nie gorsze niż 3 najlepsze dla zadanych danych algorytmy. Miarą jakości dla danych syntetycznych była średnia arytmetyczna z miary określającej pokrycie uzyskanych bi-klastrów znalezionych w zbiorze bi-klastrów oczekiwanych oraz miary określającej pokrycie bi-klastrów oczekiwanych w zbiorze bi-klastrów znalezionych.

Zaproponowane metody wykazały również że mogą poprawiać wyniki dla danych rzeczywistych. W tym celu przeanalizowano dwa zupełnie różne zbiory danych dostępne w literaturze. Wykazano że podejście znacząco poprawia jakość otrzymywanych bi-klastrów.

Powyżej opisane przeprowadzone badania obliczeniowe na danych symulowanych i rzeczywistych uzasadniają sformułowane w pracy tezy.

Oryginalną wartością dodaną omawianej rozprawy są:

- Wypracowanie miar określających podobieństwo bi-klastrów,
- Metodologia łączenia bi-klastrów oparta na uogólnionym algorytmie węgierskim,
- Meta-algorytm bi-klasteryzacji łączący wyniki różnorodnych metod,
- Dostępne publicznie oprogramowanie umożliwiające opisaną analizę we własnym zakresie.

6. Publikacje autora

Distant Analysis of the GENEPI-ENTB Databank – System Overview
Paweł Foszner, Aleksandra Gruca, and Joanna Polańska
Computer Networks, 17th Conference, CN 2010, Ustroń, Poland, June 15-19, 2010. Proceedings, pp 245-252, Volume: 79, Series ISSN: 1865-0929

Classifier Builder for DNA microarray data
Paweł Foszner, Michał Zientek
Plakat, III Zjazd Polskiego Towarzystwa Bioinformatycznego połączony z 8. Warsztatami z Bioinformatyki dla Doktorantów, 1–3 października 2010

Efficient reannotation system for verifying genomic targets of DNA microarray probes
Paweł Foszner, Roman Jaksik, Aleksandra Gruca, Joanna Polańska, Andrzej Polański
8th European Conference on Mathematical and Theoretical Biology, and Annual Meeting of The Society for Mathematical Biology, Kraków, June 28 – July 2, 2011

Improving functional coherence of gene signatures by using Gene Ontology terms
Michał Zientek, Paweł Foszner, Andrzej Polański
8th European Conference on Mathematical and Theoretical Biology, and Annual Meeting of The Society for Mathematical Biology, Kraków, June 28 – July 2, 2011

Efficient system for clustering of dynamic document database
Paweł Foszner, Aleksandra Gruca, Andrzej Polański
CDVE2011 - The 8th International Conference on Cooperative Design, Visualization and Engineering, Hong-Kong (Sep 11 - Sep 14 2011). Lectures Notes in Computer Science 6874, p.186 -189, Springer-Verlag Berlin, Heidelberg ©2011, ISBN: 978-3-642-23733-1

Efficient Algorithm for Microarray Probes Re-annotation
Paweł Foszner, Aleksandra Gruca, Andrzej Polański, Michal Marczyk, Roman Jaksik, Joanna Polańska
3rd International Conference on Computational Collective Intelligence - Technologies and Applications 21-23 September 2011, Gdynia, Poland. ICCCI'11 Proceedings of the Third international conference on Computational collective intelligence: technologies and applications - Volume Part II. Lectures Notes in Computer Science 6923, p.281-289, ISBN:978-3-642-23937-3

Comparisons of biclustering algorithms
Paweł Foszner, Aleksanrda Gruca, Andrzej Polański
Plakat, IV Zjazd Polskiego Towarzystwa Bioinformatycznego połączony z 9. Warsztatami z Bioinformatyki dla Doktorantów, Kraków, 30września–2 października 2011

Using functional coherence of gene signatures in DNA microarray data classification
Michał Zientek, Paweł Foszner, Andrzej Polański
IV Zjazd Polskiego Towarzystwa Bioinformatycznego połączony z 9. Warsztatami z Bioinformatyki dla Doktorantów, Kraków, 30września–2 października 2011

Using Gene Ontology data to improve quality of DNA microarray classification
Michał Zientek, Paweł Foszner, Andrzej Polański
V Zjazd Polskiego Towarzystwa Bioinformatycznego połączony z 10. Warsztatami z Bioinformatyki dla Doktorantów, Gdańsk, 24 – 26 maja 2012

Distributed system for computing bi-clustering algorithms
Paweł Foszner, Aleksanrda Gruca, Andrzej Polański
V Zjazd Polskiego Towarzystwa Bioinformatycznego połączony z 10. Warsztatami z Bioinformatyki dla Doktorantów, Gdańsk, 24 – 26 maja 2012

Bi-clustering – algorithms and applications
Paweł Foszner, Aleksanrda Gruca, Andrzej Polański
VI Zjazd Polskiego Towarzystwa Bioinformatycznego połączony z 10. Warsztatami z Bioinformatyki dla Doktorantów, Wrocław, 9 – 12 września 2013

Bibliografia

[1] J. N. Morgan i J. A. Sonquist, „Problems in the analysis of the survey data, and a proposal," *JAm Stat Assoc,* pp. 415-434, 1963.

[2] J. N. Hartigan, „Direct clustering of a data matrix," *JAm Stat Assoc,* pp. 123-129, 1972.

[3] B. Mirkin, „Mathematical Classification and Clustering," *Dordrecht: Kluwer.,* 1996.

[4] D. Lee i S. Seung, „Algorithms for Non-negative Matrix Factorization," *Advances in neural information processing systems,* pp. 556-562, 2000.

[5] G. Li, Q. Ma, H. Tang, A. H. Paterson i Y. Xu, „QUBIC: a qualitative biclustering algorithm for analyses of gene expression data.," *Nucleic Acids Res.,* 2009.

[6] A. Prelic, S. Bleuler i P. Zimmermann, „A systematic comparison and evaluation of biclustering methods for gene expression data," *Bioinformatics,* p. 1122–9, 2006;.

[7] G. Getz, E. Levine i E. Domany, „Coupled two-way clustering analysis of gene microarray data," *In Proceedings of the Natural Academy of Sciences,* p. 12079–12084, 2000.

[8] I. GmbH. [Online]. Available: http://ilnumerics.net/.

[9] Y. Cheng i G. M. Church, „Biclustering of expression data," *In Proc. ISMB'00,* pp. 93-103, 2000.