Akademia Górniczo-Hutnicza im. Stanisława Staszica w Krakowie

Wydział Elektrotechniki, Automatyki, Informatyki i Inżynierii Biomedycznej

AUTOREFERAT
ROZPRAWY DOKTORSKIEJ

Aggregated biclustering algorithms and their efficient implementation

mgr inż. Patryk Orzechowski

Promotor:
dr hab. inż. Krzysztof Boryczko, prof. AGH

Kraków, 2014

1 Wprowadzenie

Biklasteryzacja (ang. *biclustering*) to wywodząca się z klasteryzacji (ang. *clustering*), będącą jedną z metod rozpoznawania wzorców (ang. pattern recognition), nowoczesna technika eksploracji danych (ang. *data mining*), która znajduje szerokie zastosowanie w bioinformatyce (zwłaszcza w badaniu ekspresji genów, przy wykorzystaniu mikromacierzy DNA), jak również w analizie m.in. sieci społecznościowych, zachowań klientów czy tekstów. W odróżnieniu od klasycznych algorytmów klasteryzacji, w biklasteryzacji klasyfikacja polega na równoczesnym uwzględnieniu zarówno wzorców, jak i ich atrybutów (stąd przedrostek "bi"). Umożliwia to identyfikację podobieństw w skali lokalnej (tj. podzbiorów podobnych do siebie wzorców dla wybranych atrybutów), a nie tylko globalnej (klasyfikację wzorców na podstawie wszystkich ich atrybutów).

Analiza istniejących algorytmów pozwala stwierdzić, że *biclustering* jest zadaniem NP-trudnym. Dlatego też, zwłaszcza dla danych dużych rozmiarów, konieczne staje się zaproponowanie metod heurystycznych, których celem jest znalezienie rozwiązań suboptymalnych.

2 Cel i zakres rozprawy

Podstawowym celem rozprawy jest skonstruowanie nowoczesnych algorytmów biklasteryzacji, nazywanych „zagregowanymi", oraz zaproponowanie ich efektywnej implementacji. Poprzez „zagregowane" algorytmy, rozumiane są takie algorytmy, których struktura została zainspirowana istniejącymi rozwiązaniami. Można bowiem wyróżnić takie aspekty poszczególnych algorytmów, które determinują potencjał i siłę danej metody (np. jego koncepcja, definicja subproblemu, metryka podobieństwa, sposób przetwarzania danych czy też reprezentacja danych). Dokonując selekcji, a następnie łącząc poszczególne elementy z wybranych algorytmów, można stworzyć nową klasę algorytmów, której przedstawiciele będą rozszerzać właściwości istniejących rozwiązań.

Dla postawionego celu ogólnego zdefiniowano kilka celów szczegółowych. Pierwszy z nich polega na dokonaniu systematycznej klasyfikacji algorytmów biklasteryzacji uwzględniającej najważniejsze ich cechy od strony informatycznej, takie jak złożoność obliczeniową oraz metodologię uzyskiwania wyników. Drugim celem jest zaproponowanie zagregowanych algorytmów biklasteryzacji oraz ich implementacji. Efektywność metod zostanie zbadana z wykorzystaniem ogólnodostępnych zbiorów danych. Trzecim celem jest analiza możliwości skrócenia czasu obliczeń poprzez zaproponowanie implementacji dostosowanej do nowoczesnych masywnie równoległych architektur, z wykorzystaniem standardu OpenMP lub dla procesorów graficznych w oparciu o język OpenCL.

Proponuje się następującą tezę doktoratu: **Istnieje możliwość zaproponowania efektywnych algorytmów biklasteryzacji, które będą łączyły w sobie wybrane cechy istniejących algorytmów w taki sposób, iż poprawie ulegnie jakość uzyskanych rezultatów.**

3 Najważniejsze wyniki

W części tej zaprezentowano najważniejsze osiągnięcia niniejszej pracy. Opisano pokrótce proponowane zagregowane algorytmy biklasteryzacji, a następnie przedstawiono proponowane klasyfikacje algorytmów biklasteryzacji oraz biklastrów. Przedstawiono także wybrane wyniki zastosowania algorytmów biklasteryzacji dla danych biologicznych oraz danych pochodzących z sieci społecznościowych.

3.1 Zagregowane algorytmy biklasteryzacji

W pracy zaproponowano następujące zagregowane algorytmy biklasteryzacji:

1. *Propagation-Based Biclustering Algorithm (PBBA)* - algorytm wykorzystujący m.in. autorską koncepcję metody preprocessingu, nazwanej *vicinity transformation*, mającą na celu zlokalizowanie wszystkich lokalnych podobieństw. Algorytm zaimplementowano w C++.

2. *Qualitative Propagation Biclustering (QPB)* - stanowi pewną modyfikację algorytmu PBBA, zawierającą dodatkowy krok preprocessingu. Implementacja w C++.

3. *Algorithm for Business Continuity Plan Pitfalls Prevention (ABCPPP)* - algorytm bazujący na PBBA, przeznaczony do wspierania zarządzania produkcji poprzez wskazywanie zagrożeń dla jej ciągłości.

4. *MiniMax with Pearson Correlation (MMPC)* - algorytm bazujący na detekcji ekstremów w poszczególnych wierszach, z wykorzystaniem współczynnika korelacji Pearsona. Implementacja sekwencyjna w C++ oraz równoległa, z wykorzystaniem OpenCL.

5. *Maximal Pearson Correlation (MPC)* - modyfikacja metody MMPC, która polega na wykrywaniu wartości maksymalnych w poszczególnych wierszach. Implementacja sekwencyjna w C++ oraz równoległa dla z wykorzystaniem OpenCL.

6. *MiniMax with Spearman Correlation (MMSC)* - modyfikacja metody MMPC, która zamiast korelacji Pearsona wykorzystuje współczynnik korelacji rang Spearmana.

3.2 Klasyfikacja algorytmów biklasteryzacji

W pracy dokonano klasyfikacji algorytmów biklasteryzacji w zależności od użytej w nich miary podobieństwa. Wyróżniono następujące grupy miar, które są używane w algorytmach biklasteryzacji [2]:

- **miary bazujące na odległości**, wykorzystujące popularne metryki odległościowe (np. odległość euklidesowa, Manhattan, Czebyszewa)

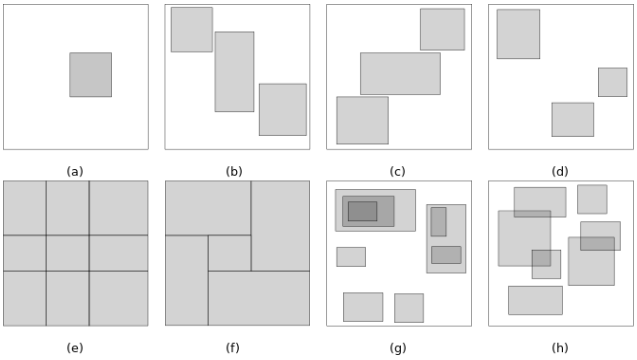
- **miary jakościowe**, zliczające najczęściej występujące elementy (np. dodatnie, ujemne) lub liczbę niezgodności pomiędzy dwoma zbiorami,

- **miary nie bazujące na korelacji** (np. Mean Square Residue (MSR)),
- **miary wykorzystujące korelację**: parametryczną (np. współczynnik korelacji Pearsona) lub nieparametryczną (np. współczynnik korelacji rang Spearmana), oraz
- **inne miary (m.in. mieszane)**.

3.3 Klasyfikacja biklastrów

Jednym z efektów pracy doktorskiej jest zaproponowanie nowatorskiej klasyfikacji struktur biklastrów, która bazuje na [1]. Wyróżniamy następujące klasy (można zauważyć, że klasy te reprezentują relację generalizacji):

- **biklastry rozłączne** (ang. *exclusive biclusters*), w której to klasie każdy wiersz (kolumna lub jedno i drugie) może zostać przypisana wyłącznie do pojedynczego biklastra (np. pojedynczy biklaster) - Rysunek 1a-d,
- **biklastry nie nakładające się** (ang. *disjoint (non-overlapping) biclusters*), która obejmuje takie biklastry, które nie współdzielą z innymi biklastrami żadnego wiersza ani kolumny (np. szachownica)- Rysunek 1e-f,
- **biklastry zawierające się** (ang. *inclusive biclusters*), które są albo rozłączne, albo całkowicie zawarte w pozostałych biklastrach (np. dendrogram lub drzewo)- Rysunek 1g,
- **biklastry arbitralne** (ang. *arbitrary biclusters*), obejmująca wszystkie przypadki arbitralnego położenia biklastrów, z możliwością pokrywania się - Rysunek 1h.



Rysunek 1: Różne struktury biklastrów: (a) pojedynczy biklaster, (b) biklastry rozłączne (kolumny), (c) rozłączne (wiersze), (d) rozłączne (kolumny i wiersze), (e-f) nie nakładające się (g) zawierające się, (h) arbitralne.

3.4 Zbiory biologiczne

W ramach pracy przedstawiono metodologię eksperymentu polegającego na analizie jakości biklasteryzacji dla zbiorów GEO Series (GSE) pochodzących z eksperymentów mikromacierzowych. Celem testu była weryfikacja możliwości zarekomendowania określonej metody preprocessingu dla danego algorytmu biklasteryzacji. Analizie porównawczej poddano kilka najpopularniejszych algorytmów biklasteryzacji oraz wybrane autorskie, które zostały przedstawione w rozprawie. Zaproponowano dwa współczynniki oceny jakości otrzymanych wyników: precyzję (ang. *precision*), która określa najbardziej wzbogaconą wartość najlepszego spośród biklastrów, oraz różnorodność (ang. *diversity*), determinującą zdolność algorytmu do wykrywania różnorodnych GO-wyrażeń w zbiorze. Wyniki pozwoliły określić zależność skuteczności wybranych algorytmów biklasteryzacji od zastosowanej metody preprocessingu.

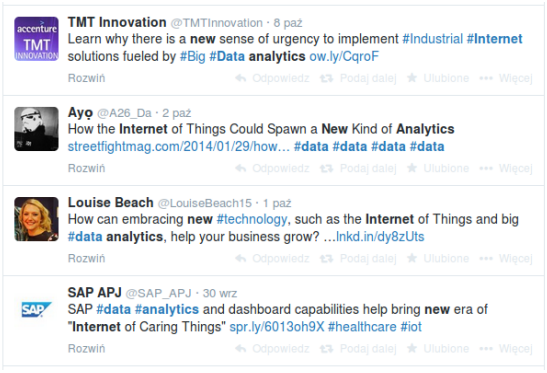
3.5 Twitter

Analizie poddano dane pobrane z Twittera zawierające słowo kluczowe "#data". Przeanalizowano w sumie 10000 tweetów w celu określenia, które słowa najczęściej występują razem w poszczególnych tweetach. Wyniki biklasteryzacji dla tak otrzymanego zbioru zaprezentowano w Tablicy 1.

Liczba wystąpień	Słowa w tym samym tweecie
1585	analytics internet new
1478	analytics customer new
1470	customer internet new
1469	analytics customer internet
1467	analytics customer internet new

Tablica 1: Biklastry dla danych pochodzących z Twittera - najpopularniejsze słowa występujące razem ze słowem #data w tweetach oraz retweetach.

Otrzymane wyniki zweryfikowano wyszukując tweety, które zawierały najpopularniejsze spośród haseł. Jak widać na Rysunku 2, biklasteryzacja potwierdziła swoją skuteczność w detekcji lokalnych podobieństw dla zbiorów tekstowych.



Rysunek 2: Przykładowe tweety, w których użyto słów „analytics” „internet” „new” oraz tagu „#data”.

4 Podsumowanie

Do najważniejszych oryginalnych rezultatów pracy zaliczyć można:

1. Przedstawienie koncepcji oraz zaproponowanie kilku algorytmów zagregowanych.
2. Implementacja algorytmu dedykowanego dla architektur masywnie równoległych z wykorzystaniem OpenCL.
3. Opracowanie klasyfikacji algorytmów biklasteryzacji bazującej na typach wykorzystywanych w nich miar.
4. Zaproponowanie zunifikowanej klasyfikacji struktur biklastrów.
5. Zaproponowanie metodologii analizy zbiorów GEO Series (GSE) oraz wskaźników oceny jakości biklasteryzacji.
6. Analizę wpływu wyboru metody preprocessingu dla zbiorów GEO Series (GSE) na jakość uzyskanych wyników.
7. Zastosowanie biklasteryzacji w nowatorskich obszarach, takich jak biklasteryzacja danych produkcyjnych (ALMM) oraz tweetów z Twittera.

5 Perspektywy kontynuacji

Istnieje szereg płaszczyzn, na których praca może być (i będzie) kontynuowana. W pierwszej kolejności planowane jest poszerzenie przeprowadzonych testów o inne zbiory danych, w celu dodatkowego potwierdzenia postawionych hipotez. Rozważa się także uwzględnienie innych algorytmów biklasteryzacji w tej analizie. W dalszej kolejności rozważa się zaproponowanie oraz implementację algorytmu biklasteryzacji dedykowanego dla architektury Intel MIC oraz dla zintegrowanych układów Accelerated Processing Unit (APU), jak również opracowanie koncepcji wizualizacji rezultatów biklasteryzacji, która będzie bazowała na Skalowaniu Wielowymiarowym (Multi-Dimensional Scaling - MDS). Przewiduje się także opracowanie metody typu ensemble lub nowych metod zagregowanych. Przedstawione problemy badawcze w dużej mierze będą realizowane w ramach grantu finansowanego ze środków Narodowego Centrum Nauki (NCN).

Rozprawa doktorska jest finansowana ze środków Narodowego Centrum Nauki (NCN), grant nr 2013/11/N/ST6/03204. Praca została wykonana z wykorzystaniem Infrastruktury PL-Grid.

Bibliografia

[1] S. C. Madeira and A. L. Oliveira. Biclustering algorithms for biological data analysis: a survey. *Computational Biology and Bioinformatics, IEEE/ACM Transactions on*, 1(1):24–45, 2004. 3

[2] P. Orzechowski. Proximity measures and results validation in biclustering – A survey. In Leszek Rutkowski, Marcin Korytkowski, Rafał Scherer, Ryszard Tadeusiewicz, Lotfi A. Zadeh, and Jacek M. Zurada, editors, *Artificial Intelligence and Soft Computing*, volume 7895 of *Lecture Notes in Computer Science*, pages 206–217. Springer Berlin Heidelberg, 2013. 2