Zadania projektowe z przedmiotu MED **(semestr zimowy 2023/2024)**

Prowadzący Grzegorz Protaziuk, email. grzegorz.protaziuk@pw.edu.pl
Konsultacje: czwartek w godz. 12.15-14.00, możliwość konsultacji w trybie zdalnym (teams) po uprzednim umówieniu się (np. mejlowo).
Platforma LeON kurs: https://leon.pw.edu.pl/course/view.php?id=1640

Wymagania

Cel: Zbadanie własności zaimplementowanego przez siebie algorytmu:

- czas wykonania dla rożnych wartości parametrów wejściowych,
- wpływ zmiany wartości parametrów na uzyskiwane wyniki,
- skalowalność algorytmu względem rozmiaru danych (liczby rekordów oraz liczby atrybutów),
- oraz cele szczegółowe, jeśli są podane w opisie zadania.

Testy należy przeprowadzić dla kilku zbiorów danych

Środowisko deweloperskie

Wybór języka programowania oraz narzędzi i bibliotek należy do wykonawcy zadania, ograniczeniem jest tu legalność wykorzystania wybranych narzędzi/bibliotek do realizacji projektów studenckich.
Preferowane są popularne języki oprogramowania.

Grupy

Zadania są pomyślane jako jednoosobowe (chyba, że wprost wskazano inaczej), jednak można realizować je w grupie maksymalnie dwuosobowej. W takim przypadku należy ustalić z prowadzącym zakres zadania projektowego.

Dokumentacja

Dokumentacja powinna zawierać:

- opis zadania – przyjęte założenia,
- opis implementacji – najważniejsze elementy projektu i implementacji
- opis przeprowadzonych eksperymentów: cel, zbiory danych, sposób wykonania
- uzyskane wyniki eksperymentów oraz wnioski
- opis formatu danych wejściowych i wyjściowych
- instrukcja użycia oprogramowania: specyfikacja parametrów wejściowych, sposób instalacji/uruchomienia.

Zbiory danych

https://archive.ics.uci.edu/ml/datasets.php,
http://fimi.uantwerpen.be/data/,
https://github.com/deric/clustering-benchmark,
http://www.philippe-fournier-viger.com/spmf/index.php?link=datasets.php

Eksperymenty można wykonać na zbiorach danych dostępnych z innych źródeł niż podane wyżej,

Terminy:

1. 2023.12.01 - wybór tematu.
2. 2023.12.21 - prezentacja kluczowych elementów rozwiązania, w tym zakresu projektu
3. 2024.01.19 – oddanie realizacji zadania.

Procedura wyboru tematu projektu do realizacji:

1. Należy przesłać e-mailem (z frazą MED. w temacie) preferencje (3 tematy) do prowadzącego (grzegorz.protaziuk@pw.edu.pl) do 01.12.2023 r. włącznie (zapisy na temat będą realizowane wg kolejności zgłoszeń i preferencji) – z wykorzystaniem konta/adresu politechnicznego.
2. Osoby zgłaszające grupę powinny wysłać ten sam e-mail do wiadomości wszystkich pozostałych członków zespołu.
3. Osoby, które chcą realizować własny temat muszą uzgodnić cel i zakres projektu z prowadzącym do dnia 01.12.2023 r.
4. Osoby, które nie prześlą preferencji w wyznaczonym terminie zostaną przypisane do tematu projektu arbitralnie wybranego przez prowadzącego.

Prezentacja kluczowych elementów rozwiązania:

Kluczowe elementy rozwiązania należy przedstawić osobiście do 17.05.2023 r. Przedstawienie obejmuje:

- przedstawienie zakresu, idei rozwiązania oraz planów testów w formie prezentacji (np. w formacie .ppt);
- omówienie planowanego rozwiązania.

Przed prezentacją kluczowych elementów projektu należy uzgodnić z prowadzącym dokładny termin.

Oddanie projektu:

Projekt należy oddać osobiście do 19.01.2024 r.
Oddanie projektu obejmuje:

1. prezentację pokazującą główne zagadnienia związane z realizowanym projektem – należy przygotować prezentację (np. w formacie .ppt);
2. pokaz działania oprogramowania
3. rozmowę dotyczącą uzyskanych wyników i wniosków.

Przed oddaniem projektu należy uzgodnić z prowadzącym dokładny termin oddania oraz przesłać/wgrać odpowiednio wcześniej (absolutne minimum to jeden dzień roboczy przed oddaniem) dokumentację projektu.

Prezentacja kluczowych elementów rozwiązania oraz oddanie projektu może być zrealizowana w trybie stacjonarnym lub zdalnym. Dla trybu zdalnego wymagane jest połączenie z wizją.

Ocenianie

Projekt jest oceniany na ocenę w standardowej skali (2-5). Ocenie podlega końcowa dokumentacja projektu, jednak brak prezentacji kluczowych elementów rozwiązania w podanym terminie powoduje obniżenie końcowej oceny o 0,5.

Tematy zadań

0. Temat własny – wymaga uzgodnienia

1. Reguły asocjacyjne

1 Implementacja algorytmu do odkrywanie reguł asocjacyjnych (Apriori, Eclat, FP-Growth, …).
2 Implementacja jeden z algorytmów opisany w artykułach:

[1] Thanh-Long Nguyen, Bay Vo, Vaclav Snasel, Efficient algorithms for mining colossal patterns in high dimensional databases, 2017, Knowledge-Based Systems.

[2] Implementacja algorytmu opisanego w artykule: Ezeife, C.I., Su, Y. (2002). Mining Incremental Association Rules with Generalized FP-Tree. In: Cohen, R., Spencer, B. (eds) Advances in Artificial Intelligence. Canadian AI 2002. Lecture Notes in Computer. Grupa 2 osobowa.

[3] Jiawei Han and Yongjian Fu, "Mining multiple-level association rules in large databases," in IEEE Transactions on Knowledge and Data Engineering, vol. 11, no. 5, pp. 798-805, Sept.-Oct. 1999, doi: 10.1109/69.806937.

2. Równoległe algorytmy wykrywania reguł asocjacyjnych

Implementacja jeden z algorytmów opisany w artykułach:

[1] R. Agrawal and J. Shafer, "Parallel Mining of Association Rules, IEEE Transactions On Knowledge And Data Engineering, Vol 8, No 6, December 1996

[2] Mohammed J. Zaki, Srinivasan Parthasarathy, Wei Li, A Localized Algorithm for Parallel Association Mining, 9th Annual ACM Symposium on Parallel Algorithms and Architectures (SPAA), pp 321-330, Newport, Rhode Island, June 22-25, 1997.

3. Grupowanie

Implementacja algorytmu powinna umożliwiać grupowanie obiektów opisanych wieloma atrybutami, zarówno numerycznymi jak i nominalnymi. Należy porównać ocenić przydatność algorytmu do grupowania obiektów opisanych atrybutami: 1) numerycznymi, 2) nominalnymi i numerycznymi.

3.1 Implementacja algorytmu CURE (Clustering Using Representative).
Literatura: Sudipto Guha, Rajeev Rastogi, Kyuscok Shim, CURE: An Efficient Clustering Algorithm for Large Databases, 1998 (citeseer.ist.psu.edu)

3.2 Implementacja algorytmu Clarans
Literatura: Raymond T.Ng, J.Han: CLARANS: A method for clustering objects for spatial data mining 2002 IEEE Transactions on Knowledge and Data Engineering

3.3 Implementacja algorytmu NBC (Neighborhood-Based Clustering algorithm)
Listeratura: S. Zhou, Y. Zhao, J. Guan, and J. Huang, “A Neighborhood-Based Clustering Algorithm,” in Advances in Knowledge Discovery and Data Mining

4. Wzorce sekwencyjne

Implementacja algorytmu do odkrywanie częsty wzorców sekwencyjnych: PrefixSpan

Literatura: Pei, J., Han, J., Mortazavi-Asl, B., Pinto, H., Chen, Q., Dayal, U., Hsu, M., “Mining Sequential Patterns by Pattern-Growth: The PrefixSpan Approach”, IEEE Trans. Knowledge and Data Engineering,

5. Reguły epizodyczne

Zaimplementować algorytm opisany w wybranym artykule:

1. N. Meger, C. Rigotti: Constraint-Based Mining of Episode Rules and Optimal Window Sizes PKDD'04, Pisa, Italy.
2. Gemma Casas-Garriga: Discovering Unbounded Episodes in Sequential Data, PKDD 2003

6. Indukcja reguł

1. Implementacja algorytmu AQ – Cervone, Guido & Franzese, Pasquale & Keesee, Allen. (2010). Algorithm quasi-optimal (AQ) learning (http://geoinf.psu.edu/publications/2010_WIRES_AQLearning_Cervone.pdf, również w P. Cichosz “Systemy uczące się” WNT