1. OPIS PROJEKTU DOKTOSKIEGO (4000 znaków max., łącznie z celami i planem pracy, do umieszczenia na stronie internetowej Szkoły)

Tytuł projektu:

Opracowanie i analiza nowych algorytmów klasteryzacji

1.1. Cele projektu

Celem zasadniczym projektu jest opracowanie nowych algorytmów klasteryzacji.
Planujemy również prace nad przyśpieszaniem istniejących algorytmów, czyli taką ich zmianę aby zredukować złożoność obliczeniową.
Wszystkie nowe lub modyfikowane algorytmy będą podlegały właściwej analizie jakościowej i złożoności.
Innym celem jest również stworzenie wersji zrównoleglonych algorytmów klasteryzacji, co jest ważne szczególnie dzisiaj, gdy mamy do czynienia z architekturami wieloprocesorowymi czy kartami GPU.

1.2. Ogólna charakterystyka projektu

Można powiedzieć, że do tej pory powstało sporo algorytmów klasteryzacji, ale wiele z nich charakteryzuje się nieciekawą złożonością (są wolne) lub nie są w stanie ekstrahować klastrów o złożonych strukturach. W ramach tego projektu planuję konstrukcję nowych algorytmów, które będą potrafiły konstruować klastery o strukturach wykraczających poza to co było realne do tej pory. Warto tu zwrócić uwagę, iż algorytmy klasteryzacji odgrywają ogromną wagę w przeróżnych zastosowaniach. Dla przykładu ostatnio są wręcz niezbędnym etapem bardzo wyrafinowanej analizy bioinformatycznej, której materiał DNA jest pobierany masowo z wybranego narządu.
Dlatego planowane jest stworzenie przede wszystkim nowych algorytmów, które będą mogły sprostać potrzebie tworzenia klastrów o strukturach znacznie bardziej wyrafinowanych niż miało to miejsce do tej pory.
Ograniczenia jakie mają obecne algorytmy może bezpośrednio uniemożliwić osiągnięcie oczekiwanych rezultatów choćby we wspomnianych zastosowaniach bioinformatycznych realizowanych choćby przez kolegę z Cincinnati.
Chcemy również opracować zmodyfikowane wersje ciekawych algorytmów tak, aby zredukować istotnie ich złożoność obliczeniową.
Zasadniczo wszystkie algorytmy, które chcemy stworzyć mają być wyłącznie algorytmami o niskiej złożoności (O(n log n)), bo tylko wtedy można oczekiwać, że przydadzą się w

obliczeniach na dużych danych. Warto zauważyć, iż choćby w ostatnich latach zajmowaliśmy się właśnie tworzeniem wyłącznie bardzo wydajnych algorytmów, więc nasze doświadczenie w takim projektowaniu algorytmów jest duże.
W planach mamy także analizę tworzonych lub modyfikowanych algorytmów od strony złożoności a także jakościowej.

Innym ważnym aspektem jest stworzenie wersji zrównoleglonych nowych lub zmodyfikowanych algorytmów. Można wręcz powiedzieć, że ich niestworzenie często bywa nieroztropne ponieważ przy dzisiejszych wielokorowych systemach komputerowych czy kartach GPU można uzyskać wielokrotne przyśpieszenie. Oczywiście jeśli tylko dany algorytm da się zrównoleglić lub zrównolegleniu podlegnie jego zmodyfikowana wersja. Dlatego planujemy także stworzenie wersji równoległych szczególnie nowoopracowywanych algorytmów.

1.3. Plan pracy

- zapoznanie się z wybranymi najnowszymi wersjami algorytmów klasteryzacji
- analiza algorytmów i struktur danych do znajdowania najbliższych sąsiadów najnowszymi metodami
- analiza porównawcza algorytmów i struktur danych do wyznaczania ANN
- przyspieszenie algorytmu Cameleon 2
- tworzenie nowych algorytmów klasteryzacji
- analiza łączenia t-SNE z nowymi algorytmami
- zrównoleglenie wybranych algorytmów
- analiza danych bioinfmormatycznych z użyciem nowych algorytmów klasteryzacji

[powyższe zadania nie planuje się wykonywać w bieżącej kolejności]

1.4. Literatura *(max. 10 pozycji/sugestia lektury dla kandydatów)*

Norbert Jankowski. “Revdbscan and Flexscan—O(n log n) clustering algorithms”. In: *Neural Information Processing*. Ed. by Teddy Mantoro et al. Cham: Springer International Publishing, 2021, pp. 642–650

M. Orliński and N. Jankowski. “Fast t-SNE algorithm with forest of balanced LSH trees and hybrid computation of repulsive forces”. In: *Knowledge-Based Systems* 206 (2020), pp. 1–

16.

Efficient and robust approximate nearest neighbor search using Hierarchical Navigable Small World graphs, Yu. A. Malkov, D. A. Yashunin, 2018, arXiv:1603.09320

Chameleon 2: An Improved Graph-Based Clustering Algorithm, Tomas Barton, Tomas Bruna and Pavel Kordik, ACM Transactions on Knowledge Discovery from Data, Volume 13 Issue 1 Article No.: 10pp 1–27

1.5. Wymagana wstępna wiedza i umiejętności kandydata/tki na doktoranta/kę

Ukończone zajęcia z uczenia maszynowego lub sieci neuronowych na poziomie studiów magisterskich z wysokim wynikiem.
Znajomość złożonych struktur danych, wynikająca np. z ukończenia zajęć Algorytmu II na kierunku Informatyka Stosowana stopień II z wysokim wynikiem. Płynna znajomość środowiska .Net/C#, C++. Umiejętność tworzenia programów zrównoleglonych.

1.6. Oczekiwany rozwój wiedzy i umiejętności kandydata/tki na doktoranta/kę

Doktorant będzie musiał rozwinąć zakres wiedzy tak, aby móc tworzyć algorytmy i analizować algorytmy a nie tylko zrozumieć jak działają algorytmy stworzone przez inne osoby w zakresie metod uczenia maszynowego.