

Katedra Grafiki, Wizji Komputerowej i Systemów Cyfrowych

Politechniki Śląskiej

Opinia

O pracy doktorskiej mgr Mateusza Kani, „Data clustering with mixtures of multidimensional distributions”

Mgr Mateusz Kania ukończył studia magisterskie na Uniwersytecie Śląskim na kierunku biotechnologia. Od 2019 roku, w Politechnice Śląskiej, prowadził badania związane z analizą danych sekwencjonowania genomowego tkanek nowotworowych, wspierane posiadanym przez nas dostępem do bazy TCGA. Analizował zależności pomiędzy ścieżkami genowymi a ekspresją genów. Opanował także metodologię analizy danych sekwencjonowania pojedynczych komórek. Na potrzeby budowania potoków analiz dla tych danych rozwijał i implementował różnorodne techniki grupowania nienadzorowanego. W szczególności opracował oryginalne implementacje algorytmów grupowania bazujące na modelach wielowymiarowych mieszanin rozkładów prawdopodobieństwa. Opracował oryginalne metody optymalizacji wiarygodności modeli wielowymiarowych mieszanin, w ramach schematu EM, umożliwiające analizy danych o wymiarach tysięcy lub nawet dziesiątków tysięcy składowych wektora danych. W swojej pracy naukowej wykazuje się bardzo dużym zaangażowaniem. Rozwinął efektywny dobry warsztat programistyczny. Potrafi konstruować funkcjonalne systemy oprogramowania przede wszystkim w środowisku R. Stale pracuje nad opanowaniem i rozwinięciem różnych elementów warsztatu modelowania matematycznego danych.

Opiniowana praca doktorska zawiera opis i podsumowanie uzyskanych przez niego wyników w zakresie konstrukcji i dopasowania do danych modeli wielowymiarowych mieszanin, a także bardzo szerokie porównania tych modeli do innych algorytmów nienadzorowanego grupowania. Przy konstrukcji algorytmów grupowania bazujących na mieszaninach rozkładów wielowymiarowych doktorant korzystał z ogólnego schematu iteracji EM (expectation maximization), jednak musiał dokonać dopasowania struktury algorytmu do postaci konkretnych rozkładów wielowymiarowych. Zastosował dwa modele mieszanin rozkładów, mieszaniną wielowymiarowych rozkładów normalnych z diagonalną macierzą kowariancji oraz mieszaniną rozkładów wielomianowych. Obok wyprowadzenia postaci iteracji opracował także odpowiednią implementację programową (w środowisku R) umożliwiającą grupowanie danych o wysokich wymiarowościach. Porównania wyników grupowania różnych algorytmów przeprowadził zarówno dla danych symulowanych jak też dla różnorodnych zbiorów danych rzeczywistych. Opracował i zaimplementował wskaźniki dla porównań różnych algorytmów. Wyniki porównań wykazały, że badane modele mieszanin stanowią wartościową alternatywę dla bardzo często stosowanych algorytmów grupowania bazujących na odległościach.

W podsumowaniu uważam, że opiniowana praca może być podstawą do ubiegania się o stopień doktora nauk w dyscyplinie Informatyka Techniczna i Telekomunikacja.

