МИНИСТЕРСТВО ОБРАЗОВАНИЯ РЕСПУБЛИКИ БЕЛАРУСЬ

БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

ФАКУЛЬТЕТ РАДИОФИЗИКИ И КОМПЬЮТЕРНЫХ ТЕХНОЛОГИЙ

Кафедра системного анализа и компьютерного моделирования

ГОРБУНОВА Анастасия Андреевна

РАЗРАБОТКА АЛГОРИТМОВ СНИЖЕНИЯ РАЗМЕРНОСТИ БОЛЬШИХ БИОМЕДИЦИНСКИХ ДАННЫХ

Аннотация к магистерской диссертации

Специальность 1–98 80 01 «Информационная безопасность»

Научный руководитель
Яцков Николай Николаевич
кандидат физико-математических наук,
заведующий кафедрой

Научный консультант
Назаров Пётр Владимирович
кандидат физико-математических наук

Минск, 2023

РЕФЕРАТ

В магистерской работе 92 страницы, в том числе 12 страниц приложений, 17 рисунков, 7 таблиц, 62 источника, 14 приложений.

Ключевые слова: алгоритмы снижения размерности данных; экспрессия генов; классификация; метилирование ДНК; микроРНК; мессенджерная РНК; критерии качества анализа; алгоритмы распараллеливания вычислений.

В магистерской работе представлены результаты разработки и сравнительного анализа алгоритмов снижения размерности данных начальной высокой размерности для решения задач биоинформатики, в частности для исследования типов раковых опухолей.

Выполнен сравнительный анализ алгоритмов методов главных и независимых компонент, стохастического вложения соседей с t-распределением, равномерного приближения и проекции, многомерного шкалирования, неотрицательного матричного разложения, автокодировщика, ансамблевых алгоритмов стэкинга, бэггинга и бустинга с целью классификации наблюдений по данным об экспрессии генов в заболевании рака груди. Анализ реализованных методов выполнен на экспериментальных данных о метилировании ДНК, экспрессии микроРНК и мессенджерной РНК молекулах, представляющих кластеры различной сложности. Проведен анализ алгоритмов распараллеливания вычислений.

На основании полученных результатов сделан вывод об эффективности работы алгоритмов снижения размерности данных и даны рекомендации по их использованию.

РЭФЕРАТ

У магістарскай працы 65 старонак, у тым ліку 12 старонак дадаткаў, 17 малюнкаў, 7 табліц, 62 крыніцы, 14 дадаткаў.

Ключавыя словы: алгарытмы зніжэння памернасці дадзеных; экспрэсія генаў; класіфікацыя; метыліраванне ДНК; мікраРНК; мэсанджарная РНК; крытэрыі якасці аналізу; алгарытмы распаралельвання вылічэнняў.

У магістарскай працы прадстаўлены вынікі распрацоўкі і параўнальнага аналізу алгарытмаў зніжэння памернасці дадзеных пачатковай высокай памернасці для рашэння задач біяінфарматыкі, у прыватнасці для даследавання тыпаў ракавых пухлін.

Выкананы параўнальны аналіз алгарытмаў метадаў галоўных і незалежных кампанентаў, стахастычнага ўкладання суседзяў з t-размеркаваннем, раўнамернага набліжэння і праекцыі, шматмернага шкаліравання, неадмоўнага матрычнага раскладання, аўтакадавальніка, ансамблевых алгарытмаў стэкінгу, бэгінгу і бустынгу з мэтай класіфікацыі назіранняў па дадзеных аб экспрэс рака грудзей. Аналіз рэалізаваных метадаў выкананы на эксперыментальных дадзеных аб метилировании ДНК, экспрэсіі микроРНК і месэнджарнай РНК малекулах, якія прадстаўляюць кластары рознай складанасці. Праведзены параўнальны аналіз алгарытмаў распаралельвання вылічэнняў.

На падставе атрыманых вынікаў зроблены высновы аб эфектыўнасці работы алгарытмаў зніжэння памернасці дадзеных і зроблены рэкамендацыі па іх выкарыстанню.

ABSTRACT

The master's thesis contains 92 pages, including 12 pages of applications, 17 figures, 7 tables, 62 sources, 14 applications.

Key words: dimensionality reduction methods; gene expression; classification; DNA methylation; microRNA; messenger RNA; analysis quality criteria; parallel computation algorithms.

The thesis presents the results of the development and comparative analysis of algorithms for reducing the dimensionality of data of initial high dimensionality for solving problems of bioinformatics, in particular for studying types of cancerous tumors.

The comparative analysis of the algorithms of principal and independent components, t-distributed stochastic neighbor embedding, uniform approximation and projection, multidimensional scaling, non-negative matrix factorization, autoencoder, ensemble algorithms stacking, bagging and boosting ensemble algorithms was performed in order to classify observations according to data on gene expression in breast cancer. Comparative analysis of the implemented methods was performed on three modalities: DNA methylation, microRNA and messenger RNA expression, which representing clusters of varying complexity. A comparative analysis of parallelization algorithms has been carried out.

Based on the obtained results, conclusions about the efficiency of dimensionality reduction methods were drawn and recommendations for their use were given.