ПРАВИТЕЛЬСТВО РОССИЙСКОЙ ФЕДЕРАЦИИ
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ
ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ
«ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»

Факультет компьютерных наук

УТВЕРЖДАЮ
Академический руководитель
образовательной программы
«Науки о данных»,

_____ С.О. Кузнецов
«___» _____ 2018 г.

Выпускная квалификационная работа

на тему
Разработка программного обеспечения, ориентированного на пользователя, для проведения кластер-анализа
по критерию наименьших квадратов

тема на английском языке
Developing a user-friendly software for least-squares clustering

по направлению подготовки 01.04.02 «Науки о данных»

Научный руководитель Профессор, НИУ ВШЭ Должность, место работы д.т.н, профессор ученая степень, ученое звание Б.Г. Миркин И.О. Фамилия Оценка Подпись, Дата	Выполнил студент группы мНоД16 ТМСС 2 курса магистратуры образовательной программы «Науки о данных» П.А. Еремейкин И.О. Фамилия Подпись, Дата

Москва 2018

Содержание

1 Введение

В настоящее время отмечается интенсивное развитие информационных технологий, появляются новые разработки, которые позволяют применять передовые программные решения в широком спектре областей. Если раньше информационные технологии были областью интересов узкого круга специалистов, то сейчас наблюдается тенденция к повсеместному распространению прикладных программных продуктов.

Современные компании вынуждены опираться на применение информационной инфраструктуры и использовать преимущества цифровых технологий для поддержания конкурентоспособности своих продуктов или услуг. В процессе эксплуатации информационных систем накапливаются массивы данных, обработка и интерпретация которых может принести компании коммерческую выгоду.

Каждый случай обработки данных как правило, требует индивидуального подхода, не существует универсальной последовательности операций для любой задачи. Поэтому, обработка данных требует больших интеллектуальных усилий от высококвалифицированных специалистов. Программные системы анализа данных призваны облегчить этот труд и предоставляют в распоряжение специалиста наиболее востребованные процедуры обработки данных. Особенно актуально применение таких систем для решения прикладных задач, которые, несмотря на свою индивидуальность, зачастую однотипны и имеют некоторые общие этапы решения. Таким образом, применяя готовый, тщательно отлаженный и документированный программный код, сокращается время на анализ данных, а также снижается вероятность возникновения ошибок.

Прогресс технических средств в области сбора и обработки информации приводит к росту размеров массивов данных, которые требуется обрабатывать для удовлетворения потребностей компаний. Поэтому растёт роль методов агрегации данных и выделения в них общих закономерностей или структур. К таким методам, в частности относятся методы кластеризации.

Под кластеризацией понимают выделение объектов из таблицы наблюдений в множества, называемые кластерами, которые объединяют наиболее сходные объекты, при этом различные объекты должны попадать в разные кластеры [**?**]. Задачи кластеризации часто встают в самых разных областях, например, при обработке изображений или биологических структур а также социальных групп [**?**].

Вероятно, наиболее популярный метод кластеризации — k-means.

2 Теоретическая часть