Министерство образования и науки Российской Федерации
Санкт-Петербургский политехнический университет Петра Великого
Институт прикладной математики и механики
Высшая школа теоретической механики

Работа допущена к защите

Директор высшей школы

_____ А.М. Кривцов

«___» _____2020 г.

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА БАКАЛАВРА

КЛАСТЕРНЫЙ АНАЛИЗ ДАННЫХ О ДВИЖЕНИИ ЧЕЛОВЕКА

по направлению 01.03.03 «Механика и математическое моделирование»
по образовательной программе
01.03.03_02 «Биомеханика и медицинская инженерия»

Выполнил
студент гр. 3630103/60201 П.Г Букреев

Руководитель
доцент, к.ф.-м.н. М.Б. Бабенков

Санкт-Петербург

2020

САНКТ-ПЕТЕРБУРГСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ ПЕТРА ВЕЛИКОГО

Институт прикладной математики и механики

Высшая школа теоретической механики

УТВЕРЖДАЮ

Директор высшей школы

А.М. Кривцов

« » 20 г.

ЗАДАНИЕ

на выполнение выпускной квалификационной работы

студенту Букрееву Павлу Геннадьевичу 3630103/60201

1. Тема работы: Кластерный анализ данных о движении человека

2. Срок сдачи студентом законченной работы: 15 июня 2020 г.

3. Исходные данные по работе: справочная литература, актуальные публикации по теме исследования

4. Содержание работы (перечень подлежащих разработке вопросов): подобрать подходящее устройство для сбора данных о движении человека. Выбрать метод обработки исходных данных и определить параметры для передачи алгоритму анализа. Сравнить результаты применения к данным различных алгоритмов машинного обучения. Установить предел допустимых шумов в исходных данных.

5. Перечень графического материала (с указанием обязательных чертежей): не предусмотрены

6. Консультанты по работе: не предусмотрены

7. Дата выдачи задания: 3 февраля 2020 г.

Руководитель ВКР М.Б. Бабенков

Задание принял к исполнению 3 февраля 2020 г.

Студент П.Г. Букреев

РЕФЕРАТ

На 38 с., 9 рисунков, 1 таблицу.

КЛАСТЕРНЫЙ АНАЛИЗ, МАШИННОЕ ОБУЧЕНИЕ, МЕТОД К-СРЕДНИХ, РАСПОЗНАВАНИЕ ЧЕЛОВЕЧЕСКОЙ ДЕЯТЕЛЬНОСТИ, БИОМЕХАНИКА

В рамках данной выпускной работы было проведено сравнение трех методов кластерного анализа данных о движении человека. Работу можно разделить на три части: поиск подходящего устройства для получения исходных данных, поиск минимально необходимого набора признаков для выделения из исходных данных, применение и сравнение результатов трех рассмотренных методов кластерного анализа.

THE ABSTRACT

38 pages, 9 pictures, 1 table

CLUSTER ANALYSIS, MACHINE LEARNING, K-MEANS CLUSTERING, HUMAN ACTIVITY RECOGNITION, BIOMECHANICS

In this work, a comparison of three different clustering methods was performed within the human activity recognition problem. The work can be divided into three parts: finding the appropriate device to record raw data, finding the minimum number of features required to be extracted from raw data, application and performance comparison of three clustering methods.

СОДЕРЖАНИЕ

ВВЕДЕНИЕ 6

ГЛАВА 1. ОБЩИЕ СВЕДЕНИЯ 9

1. Анализ движения человека 9

1.1. Медицинское применение 9

1.2. Спортивное применение 10

1.3. Повседневное использование 10

2. Биомеханика активностей 11

2.1. Биомеханика стояния 12

2.2. Биомеханика ходьбы 13

ГЛАВА 2. АНАЛИЗ ДАННЫХ 16

3. Выбор датчика для сбора данных 16

3.1. Преимущества смартфона 16

4. Исходные данные 18

4.1. Предварительная обработка данных 20

5. Кластеризация 20

5.1. Метод k-средних 21

5.2. EM-алгоритм 22

5.3. Иерархический подход 22

5.4. Теоретическое сравнение алгоритмов 23

6. Реализация алгоритмов 25

7. Обработка исходных данных в STATISTICA 28

7.1. Нормализация данных 28

7.2. Среднеквадратическое отклонение 28

7.3. Преобразование Фурье 30

8. Результаты 31

8.1. Метод k-средних 31

8.2. EM-алгоритм 32

8.3. Метод дерева групп 33

ЗАКЛЮЧЕНИЕ 35

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ 37

ВВЕДЕНИЕ

Распознавание человеческой активности играет важную роль в повседневной жизни людей благодаря своей возможности извлекать ценную нетривиальную информацию о выполняемых человекам действиях на основе данных с простых датчиков. В основном существует два различных типа распознавания человеческой активности: основанный на данных видео, и основанный на данных датчиков. Тип, основанный на видео данных анализирует изображения или видеоряд содержащие движения человека, в то время как второй тип нацелен на интерпретацию данных с различных датчиков, включающих в себя акселерометр, гироскоп, Bluetooth-устройства, датчики звука и прочие. Из-за активно развивающихся технологий датчиков и повсеместных вычислений, распознавание человеческой активности на основе данных датчиков становится все более популярным. В данной работе будет рассматриваться именно такой тип.

Кластеризация – процесс объединения схожих объектов в данных в группы. Объекты внутри одной такой группы отличаются от объектов внутри другой. Кластерный анализ является важным способом анализа, находящим применение во многих практических сферах.

На данный момент не существует строгого определения термина «активность». Однако, для всех его применений характерно понятие дробимости и делимости на более примитивные составляющие. Так, например, активность «сварить кофе» может быть разделена на составляющие: «зайти на кухню», «взять кружку» и т.д.; эти более простые активности в свою очередь также состоят из еще более примитивных. Таким образом, под активностями понимаются наборы действий, которые в свою очередь состоят из более примитивных действий.

Как правило, активность предполагается выполняемой в пределах определенного временного промежутка, который складывается из продолжительностей состоящей в ней более простых активностей. Исходя из

таких предположений, можно выделить различные типы активности: сложные активности, последовательные активности, одновременные активности и взаимосвязанные активности. Подводя итог вышесказанному, активности и действия могут быть извлечены из исходных данных с применением корректной иерархии активностей с привлечением дополнительной контекстуальной информации.

Распознавание движения может быть представлено как типичная проблема распознавания закономерностей. Традиционные подходы к текущему моменту уже достигли значительного прогресса в распознавании движения, пользуясь такими алгоритмами машинного обучения как дерево выборов, метод опорных векторов, наивный байесовский классификатор, и скрытая марковская модель. Достаточно очевидно, что в некоторых контролируемых условиях, для которых характерно малое количество помеченных данных или определенное знание размерности данных, традиционные методы полностью справляются с получением удовлетворительных результатов. Однако, в большинстве повседневных задач распознавания движения, такие методы зачастую значительно опираются на эвристические методы извлечения признаков. Также, такие подходы могут определять только примитивные классы, что ведет к снижению применимости для задач без учителя. В связи с этими ограничениями, применимость традиционных методов ограничена в смысле точности классификации и возможности обобщения модели.

Распознавание движения стремится понять поведение человека. Говоря формальным языком, предполагается что пользователь выполняет некоторую активность, принадлежащую к заранее определенному набору активностей A:

$$A = \{A_i\}_{i=1}^{m}$$

где m обозначает количество различных типов активности. Также существует последовательность отрезков данных описывающих активность, полученных с датчиков

$$\boldsymbol{s} = \{\boldsymbol{d}_1, \boldsymbol{d}_2, \ \cdots, \boldsymbol{d}_t, \ \cdots \ \boldsymbol{d}_n\}$$

где $\boldsymbol{d}_t$ обозначает сигнал датчика в момент времени t.

Целью данной работы является кластерный анализ данных о движении человека. В связи с поставленной целью в работе решаются следующие задачи:

1. подобрать подходящее устройство для сбора данных о движении человека.
2. Выбрать метод обработки исходных данных и определить параметры для передачи алгоритму анализа.
3. Сравнить результаты применения к данным различных алгоритмов машинного обучения.

ГЛАВА 1. ОБЩИЕ СВЕДЕНИЯ

1. Анализ движения человека

С недавним развитием такого феномена как интернет вещей, применение анализа движения стало значительно более распространенным. Оно используется для распознавания различных повседневных активностей и не требует значительных затрат на специализированные датчики, роль которых в современной ситуации с лихвой выполняют сравнительно доступные смартфоны, фитнес браслеты и прочие устройства для повседневного ношения. Конкретными датчиками, собирающими информацию в этих устройствах, чаще всего являются акселерометр, гироскоп и датчик сердцебиения.

Анализ движения человека по большей части используется в следующих сферах [7, стр. 1054-1058]:

1.1. Медицинское применение

Не все заболевания могут быть предотвращены, диагностированы или излечены стандартной практикой, для которой необходимо продолжительное по времени наблюдение пациента, зачастую и во время выполнения бытовых задач. В случае некоторых заболеваний, симптому могут не проявиться в течение медицинского наблюдения, что дает возможность анализу движения предоставить докторам дополнительную информацию о различных отклонениях от нормы в процессе выполнения пациентом его ежедневной рутины. Такая информация может сыграть ключевую роль в более раннем определении таких диагнозов как инсульт [15]. Анализ движения также полезен в диагностировании болезни Паркинсона, для определения симптомов и мониторинга протекания заболевания [8]. Реабилитация и наблюдение за процессом реабилитации также являются одной из областей применения распознавания движения [6, стр. 810-814].

В рамках медицинской сферы распознавание движения используется в том числе и в области диетологии [10]. Метод автоматического определения

процесса приема пищи или потребления напитков могут быть использованы для уведомления пользователей об их привычках. Схожий метод может быть применен и для уведомления курящих пользователей или как более дешевая альтернатива таким методам анализа функций сна как полисомнография [13, стр. 886-891; 4, стр. 21-28].

1.2. Спортивное применение

Задача фитнес-трекинга (или применения в спорте) заключается в отслеживании таких значений как пройденные шаги, потраченные калории и сердцебиение. Некоторые более сложные системы могут давать значительно более сложно наблюдаемых значений, таких как пройденные этажи или качество сна.

Как уже было сказано ранее, современные смартфоны содержат встроенные акселерометр, гироскоп и систему глобальной спутниковой навигации, которые в совокупности предоставляют широкому кругу пользователей отслеживать фитнесс активность. Как правило, приложения для смартфона для анализа активности выполняют упрощенную версию анализа, состоящую из выделения отдельных шагов, пройденного расстояния и скорости шага, дополняя анализ грубым делением активностей на «стояние» и «нахождение в движении». Также таким приложениям зачастую необходима дополнительная информация от пользователя для распознавания более точных активностей, таких как плавание или бег.

1.3. Повседневное использование

Распознавание активности может быть настроено для работы со сложными активностями и состояниями пользователя, такими как «на работе» или «на пути домой», что может быть использовано другими устройствами или приложениями, например, для включения режима «не беспокоить» во время работы.

Подтипом задачи анализа человеческой активности является задача определения способа передвижения, которая заключается в определении

используемого транспортного средства, будь то поезд, автомобиль, велосипед или другой вид личного транспорта. Эта информация также может быть использована для предоставления контекста другим устройствам или приложениям на смартфоне.

Среди других применений такой контекстуальной информации можно указать использование ее для саморегулирующихся систем. Примером может служить отключение некоторых функций смартфона, таких как Bluetooth или WiFi во время бега, или отключение служб геопозиции, если пользователь не передвигается. Также стоит отметить возможность снижать частоту считывания данных устройствами смартфона, если для определения параметров активности достаточно более низкого значения чем установлено по умолчанию [17].

2. Биомеханика активностей

Влияние механики на биологические системы наблюдается на многих уровнях, начиная от молекулярного или клеточного и заканчивая тканями, отдельными органами или даже целыми организмами.

На самом крупном уровне, уровне всей системы (организма), механические факторы влияют на форму и выполняемые задачи мышечно-скелетной системы. Движения человека достигается за счет сложного и хорошо скоординированного взаимодействия костей, мышц, связок и суставов в составе мышечно-скелетной системы под контролем нервной системы. Мышцы создают растягивающую силу и моменты в суставах с короткими рычагами силы для обеспечения статической и динамической устойчивости тела под воздействием гравитационной или прочей нагрузки [8; 16, стр. 181]. Любое повреждения какого-либо отдельного элемента мышечно-скелетной системы приводит к изменению механических взаимодействий и приводит к ухудшению, нестабильности или невозможности движения. Однако, из соображений этики, прямое измерение сил в человеческом теле возможно только в исключительных случаях, таких как специально изготовленные имплантаты. Дальнейшим усложнением задачи становится избыточный характер мышечно-скелетной

системы. В человеческом теле находится большее количество суставов и мускул чем необходимо для выполнения повседневных задач. Следовательно, некоторые действия могут быть выполнены различными мышечно-скелетными стратегиями. В то же время, такой механизм компенсации критически важен для преодоления организмом последствий травм или заболеваний мышечно-скелетной системы. Такая избыточность накладывает дополнительные сложности для неинвазивного определения сил.

2.1. Биомеханика стояния

Необходимое условие статического равновесия с трудом применимо для человеческого тела из-за его многосегментного строения. В случае человеческого тела, множественные межсегментные силы должны уравновешивать друг друга и иметь суммарную силу равную нулю. Для стояния человека допустима некоторая гибкость, проявляющаяся в возможности центра масс совершать вертикальное и горизонтальное движение в пределах некоторого устойчивого положения, при этом не приводя к падению человека. Для передвижения центра масс из вертикальной позиции над центром устойчивой области или из горизонтальной позиции слева или справа от нее обратно в центр области требует равенство нулю интегралу сил по времени:

По горизонтали По вертикали

$$\int F_{\text{ступни}} \mathrm{d}t = 0 \qquad \int F_{\text{ступни}} \mathrm{d}t - mgt = 0,$$

где $F_{\text{ступни}}$ – результирующая сила, действующая на поверхности ступни; m – масса тела человека.

Это простое уравнение соответствует простоте описываемого им движения тела. Люди способны выполнять множество различных задач используя свои руки, двигая их из стороны в сторону без перемещения центра масс за пределы устойчивой области.

2.2. Биомеханика ходьбы

Общей механической проблемой двуногого перемещения является создание достаточного вертикального импульса для преодоления действия гравитации, вместе с этим совершая достаточную работу мышц по перемещению отстающей ноги вперед для следующего шага. Двуногое перемещение является нетривиальной задачей не только с точки зрения выполнения человеком, но и с точки зрения инженерной попытки имитирования процесса.

Для начала, рассмотрим поведение центра масс во время ходьбы. Мы можем анализировать его движение либо с точки зрения соотношений импульсов и моментов, либо с точки зрения работы и энергии. Горизонтальное движение материальной точки должно быть вызвано наличием изначально приложенной силы F на промежутке времени t для появления импульса mv по следующему закону:

$$\int_0^t F\mathrm{d}t = m(v_t - v_0)$$

Появившийся импульс будет сохраняться до момента появления противоположно направленного. Следовательно, все последующие импульсы должны в сумме равняться нулю. Ходьба также использует движение отдельных сегментов тела, что приводит к необходимости также учитывать и момент импульса. Из-за цикличности ходьбы, кинематические соотношения в любой момент внутри цикла практически идентичны таким же соотношениям в тот же момент последующего. Таким образом, моменты импульса тоже должны в сумме равняться нулю. Те же рассуждения применимы и в случае рассмотрения энергии в строгом физическом смысле, если использовать запасание энергии для замедления движения и ее выброс для ускорения движения. В обоих случаях неизменным остается тот факт, что на и на выработку механического импульса, и на совершение механической работы тратится энергия метаболизма.

Для ходьбы человеческий организм оснащен двумя соединенными нижними конечностями, в процессе движения переносящий массу верхней части тела. Рассматривая вращательное движение только в вертикальной плоскости по направлению движения, общее описание состоит из лодыжки, колена и бедра, составляющих три степени свободы. Если для каждой точки рассмотреть два вида моментов – сгибающий и разгибающий, то получится шесть механизмов контроля. Далее для каждого из уже рассмотренных моментов можно ввести два способа – концентрическое сжатие мышц и эксцентрическое сжатие мышц, то итоговое значение механизмов контроля достигает 12. Очевидно, что управление движением может быть достигнуто множеством способов, но не все из них будут эффективными. Точная целевая функция ходьбы неизвестна. Казалось бы, минимизация затрат энергии была бы полезной для организма, что подкрепляется большей эффективностью эксцентрических сжатий мышц по сравнению с концентрическими. Но второстепенной целью может также являться минимизация возмущений головы, где находятся органы баланса и ориентации в пространстве. Кинематический и кинетический анализ ходьбы дает решение проблемы ходьбы, но не дает ответа на то, какая целевая функция используется в решении. В то же время, существуют некоторые фазы ходьбы, которые могут быть объяснены при обращении к экспериментальным работам.

Известно, что ходьба состоит из последовательных периодов опоры на одну или две ноги, и что в фазе восстановления конечность, находящаяся сзади, переносится вперед для следующего периода опоры. Для начала рассмотрим хождение по прямой с постоянной скоростью. На самом деле, такое движение не существует в реальной жизни, даже если скорость центра масс тела одинакова в одинаковые моменты ходьбы. Центр масс меняет скорость как в горизонтальном, так и в вертикальном направлении из-за прикладываемых к телу сил. Также, наличие у тела различных сегментов означает появление различных моментов импульса и прироста кинетической энергии. Это вызвано тем, что нижние конечности совершают движение, похожее на движение маятника, для того

чтобы ведущая нога находилась в подходящей точке для поддержания центра масс при его горизонтальном движении.

Ключевой проблемой является продолжение процесса хождения при условии поддержания верхней части тела в устойчивом положении на каждом шаге, в то же время создавая достаточную движущую силу в горизонтальном направлении для преодоления диссипации энергии вызванной трением ступни. Вертикально направленный импульс вызван вертикально направленной силой, которая будет то больше, то меньше чем вес тела. Но суммарный импульс всегда будет равняться mgt гравитации. Горизонтально направленный импульс будет вызван силой, которая будет иногда направлена по направлению движения с помощью мышц растяжения в лодыжке, колене и бедре, а иногда против направления движения из-за трения ноги движущейся вперед. Суммарный горизонтальный импульс будет равняться нулю при условии поддержания постоянной скорости ходьбы [2].

ГЛАВА 2. АНАЛИЗ ДАННЫХ

3. Выбор датчика для сбора данных

Ранее проведенные работы показали возможности использования нательных датчиков для распознавания движения и активностей. Несмотря на многообещающие возможности такого расположения датчиков, в действительности их использование в повседневной жизни человека не представляется разумным. Например, случаи размещения множества датчиков по всему телу и наблюдение в лабораторных условиях с трудом переносятся на использование обычным человеком. Для совершения такого переноса необходимо, чтобы используемые датчики были незаметны для носителя в повседневном использовании. На текущий момент роль такого датчика могут выполнять мобильные телефоны, и в частности смартфоны.

Наличие мобильных телефонов широко распространено в большинстве стран на сегодняшний день. С течением времени, количество мобильных телефоном, содержащих встроенные датчики и множество способов соединения и передачи собранных данных, значительно увеличивалось, что дало возможность мобильным телефонам играть роль посредника между пользователем и некими контекстуальными системами. Например, возможность использовать данные акселерометра и звука окружения пользователя, переданная через беспроводную сеть может быть использована для получения информации об окружении. На текущий момент, такой подход с легкостью осуществим при помощи коммерчески доступных смартфонов. В данной работе используются данные, полученные со смартфона Apple iPhone XS.

3.1. Преимущества смартфона

У использования смартфона как наблюдающего устройства есть следующие преимущества:

- Встроенные датчики. При условии достаточности получаемой информации, пользователю не требуется использовать какие-либо

другие внешние датчики для сбора информации. В тех случаях, когда информации со встроенных датчиков недостаточно, дополнительные датчики могут быть подключены к смартфону для расширения возможностей сбора информации. Тем самым, становится очевидной многофункциональность смартфона как основного датчика.

- Современные смартфоны уже реализуют множество контекстуально-зависящих систем. Для этих нужд они оснащены сравнительно большими вычислительными мощностями и достаточными объемами памяти. Также для них характерна возможность с легкостью хранить собранные данные и затем с той же легкостью передавать их для дальнейшего использования пользователем или какой-либо системой. В общем, смартфон может рассматриваться как портативное вычислительное устройство с широкими возможностями подключения и передачи информации.
- Для смартфона высока вероятность находиться непосредственно при пользователе в течение дня. Таким образом, смартфон становится естественным кандидатом на роль датчика, не мешающего пользователям. Из-за привычек пользования, смартфон вызывает меньший дискомфорт по сравнению с альтернативными способами установки датчиков.
- Большинство смартфонов обладают сравнительно долгим временем работы. Для большинства пользователей, в условиях нормального пользования, смартфон остается в заряженном состоянии по крайней мере в течение дня, после чего может потребоваться подзарядка. Используя ранее упомянутые способы контроля частоты датчика, сбор данных на протяжении дня является возможным.

В рамках данной работы данные собирались при помощи встроенного в смартфон акселерометра. Во избежание быстрой разрядки телефона, частота

собираемых данных была установлена на значении 20 Гц. Сбор данных осуществлялся через интерфейс приложения Accelerometer. Затем данные отправлялись в формате XML на компьютер для дальнейшей предварительной обработки. Детали предварительной обработки будут представлены в следующем разделе.

Для корректного анализа движения также важно знать, насколько сильно изменение положения смартфона относительно тела влияет на точность распознавания активности. Из ранее проведенных исследований известно, что большинство мужчин носят смартфоны в переднем кармане брюк, в то время как женщины носят их в сумках [3, стр. 483-492]. Также известно, что расположения датчика чуть ниже пояса или на поясе способствует получению более точных результатов. Поэтому, для текущей постановки эксперимента, пользователь носил смартфон в привычном для себя месте, таком как например карман брюк.

4. Исходные данные

В рамках данной работы рассматриваются 6 типов активности. Большинство этих активностей характеризуются повторяющимися движениями (шаг, восход по лестнице, спуск по лестнице), что способствует более точной работе алгоритма. Таким образом, собранные данные хорошо подходят для проведения исследования и сравнения результатов различных методов.

Данные различных активностей, полученные с триаксиального акселерометра, сохраняются для каждой оси отдельно. Ось x отвечает за горизонтальное движение, отвечающее за левую и правую сторону смартфона. Эти данные используются для получения информации о горизонтальном движении ноги носителя. Ось y определяет вертикальное движение, соответствующее верхней и нижней стороне смартфона, в то время как ось z соответствует движению по направлению в и из экрана, что может быть трактовано как движение ноги вперед и назад во время выполнения активности.

В целом, все рассмотренные виды активности, за исключением сидения и стояния, могут быть исчерпывающе описаны промежутком времени между пиками ускорений и амплитудами этих пиков.

Данные для класса активности бег характеризуются отчетливыми пиками ускорения по оси x. Также данные по этой оси являются ключевыми для отделения текущей активности от других. Промежуток времени между пиками ускорения также значительно меньше чем в других классах, например, ходьбе.

Данные по оси y для класса бег достаточно сильно отличаются от остальных классов, за исключением класса ходьбы, где оба колеблются от нулевого значения до пикового. Для данной оси рекомендуется применение фильтрации гравитационного воздействия для улучшения качества работы методов анализа.

Данные по оси z класса бег также схожи с данными по той же оси для класса ходьбы. При этом пиковые отрицательные значения в классе бег превышают по модулю пиковые отрицательные значения в классе ходьба. Это явление может быть объяснено приложением больших сил в процессе бега.

Для стационарных классов (сидение и стояние), для оси x характерны малые колебания значений. Единственное различие между классами стояния и сидения может быть отмечено в относительной разнице амплитуд по разным осям, вызванной различным положением устройства на протяжении этих активностей.

Для оси y данных классов сидения и стояния практически не отличаются и, как и в случае ускорений по оси x, характеризуется малыми колебаниями.

При рассмотрении всех трех осей в совокупности, сравнение с классами движения (бег и ходьба) показывают ключевое отличие в показаниях оси z, где отсутствует характерное колебание значений вызванных движением устройства в горизонтальном направлении.

Рассмотрим отдельно класс ходьбы: для всех трех осей характерно наличие пиковых значений, отличающихся по амплитуде и периодичности.

Для класса спуск по лестнице, при рассмотрении показаний по оси x наблюдается набор равномерно распределенных колебаний около нуля. Значения ускорения по оси y достигают пикового значения в полтора G на каждом шаге по лестнице. В то же время ускорение по оси z значительно менее выраженное чем при беге, но все же сохраняющее схожие закономерности.

4.1. Предварительная обработка данных

В данной работе на основе исходных необработанных данных высчитываются значения, наиболее корректно и явно отображающие характерные свойства отдельных активностей. Ключевым понятием для таких значений является понятие признака. Признак – это индивидуальное измеримое свойство или характеристика наблюдаемого явления. Выбор информативных, отличительных и независимых признаков является критическим шагом для эффективных алгоритмов в распознавании образов, классификации и регрессии.

В данной работе из исходных данных извлекаются следующие признаки: среднее, среднеквадратическое отклонение и корреляция. Особый интерес представляла задача минимизации количества признаков необходимых для получения удовлетворительных результатов, из чего было принято решение ограничиться использованием среднего и среднеквадратического отклонения от самих показателей ускорения, а также среднее и среднеквадратическое отклонение от подвергнутых преобразованию Фурье исходных данных. Подробнее о методах вычисления этих признаков будет сказано в последующих разделах.

5. Кластеризация

Кластеризация – процесс дробления данного набора объектов на разделенные кластеры. Это делается таким образом, чтобы объекты внутри

одной такой группы (кластера) отличались бы по своим признакам от объектов внутри другой. Для извлечения полезной информации из таких кластеров, они должны отображать реальные структуры в наборе объектов.

Определение по схожести: Кластер – набор объектов, имеющих свойство «схожести»; объекты в разных кластерах не «схожи». Одним из предлагаемых способов определения «схожести» является «набор точек в пространстве, образующих вместе область с униформным локальным свойством (напр. плотность или форма).

В общем случае выделяют пять типов алгоритмов кластеризации. Наиболее используемыми из них типами являются вероятностные алгоритмы и иерархические алгоритмы.

5.1. Метод k-средних

Метод k-средних является наиболее популярным и продуктивным алгоритмом для выявления кластеров. Метод k-средних относится к вероятностным методам анализа. Из минусов алгоритма можно отметить тот факт, что при увеличении размера набора данных, затрачиваемое время и вычислительная сложность так же увеличиваются. Более того, для данного метода характерна зависимость получаемых на выходе кластеров от случайных начальных условий выбора центроид. Некоторыми исследователями было проведено несколько попыток улучшить качество работы алгоритма k-средних [11; 14, стр. 273-309]. Основная идея метода заключается в том, чтобы классифицировать данный набор данных по k разделенным кластерам, где k – заранее определенная константа, отражающая искомое количество групп. Метод подразделяется на два этапа: на первом этапе определяются k центроид, по одной на каждый кластер. Следующий этап заключается во взятии точки из набора данных и отнесении ее к ближайшей центроиде. Для определения расстояния берется евклидово расстояние в пространстве; тем же способом определяется расстояние для каждого из кластеров, равное расстоянию между каждой точкой кластера и его центроидой.

Ниже приведён псевдокод метода k-средних:

Входные данные: D = {d_1, d_2, ... , d_n} // набор n объектов

k // Количество искомых кластеров

Выходные данные: Множество k кластеров.

Шаг 1: Случайным образом выбрать k элементов-объектов из D в качестве исходных центроид; Шаг 2: Повторять Присвоить каждому элементу d_i кластер с наименьшим расстоянием до его центроиды; Вычислить новую среднюю для каждого кластера; Пока не будут удовлетворены критерии сходимости [11; 14, стр. 273-309].

5.2. EM-алгоритм

EM-алгоритм основывается на вычислении ожидаемого значения функции правдоподобия и последующей максимизации этого значения. Данный алгоритм определяет кластер по принципу его правдоподобия. На первом этапе работы алгоритма (ожидании) элементы-объекты соотносятся с центрами. На этапе максимизации значения центров кластеров обновляются с условием максимизации функции правдоподобия, вычисленной на предыдущем этапе.

Псевдокод первого шага: оценить E(z) для каждого z, θ дано

Второй шаг: оценить θ максимизируя E(log L), для данного E(z) [где “E(logL)” с учётом случайного z ~ E(z) = p(z=1)]

В первом шаге вычисляется распределение по признакам с заданной фиксированной моделью, во втором выбираются новые параметры модели для максимизации логарифмического правдоподобия ожидаемой функции учитываемых объектов и скрытых переменных [9, стр. 367-369].

5.3. Иерархический подход

Иерархическая кластеризация является ещё одним широко используемым методом кластерного анализа. Этот метод часто применяется из-за своего широкого спектра применения. Группы образуются в результате вложенной

последовательности групп (кластеров различного порядка). Иерархическая кластеризация используется для объединения схожих объектов в крупные группы в виде дерева. Алгоритм относится к виду агломеративных методов.

Псевдокод агломеративного метода:

Шаг 1: Создать для каждой точки-объекта в исходных данных собственный кластер,

Шаг 2: Вычислить все попарные расстояния между объектами, принадлежащими пространству данных

2 повторять

Шаг 3: Найти пару ближайших друг к другу кластеров;

Шаг 4: Объединить два кластера в один новый кластер *c*;

Шаг 5: Вычислить расстояние от *c* до всех остальных кластеров;

Пока не останется только один кластер

5.4. Теоретическое сравнение алгоритмов

Метод дерева групп относится к иерархическим способам кластеризации. Иерархическая кластеризация имеет несколько преимуществ по сравнению с методом k-средних и прочими алгоритмами, основанными на разделении. Метод дерева групп реализует подсчет расстояние между группами руководствуясь критериями схожестями. Также, метод k-средних имеет в результате фиксированное количество k групп, в то время как иерархический метод позволяет проводить исследование процесса объединения с любой точностью (на любом масштабе). Некоторые исследования выдвигают теорию о большей эффективности агломеративных иерархических методов по сравнению с методом k-средних. Так же этот метод может находить группы любых наперёд не заданных форм, например, с использованием одиночной связи. В то еж время, иерархические методы имеют набор отрицательных характеристик. Временная

сложность метода k-средних линейна O(n). EM-алгоритм занимает то же количество времени что и предыдущий, при условии одинаковых входных данных. Для метода дерева групп временная сложность квадратичная. Также метод дерева групп очень чувствителен к выбросам в исходных данных [12].

Таблица 5.4. Сравнение методов кластеризации

Показатель	Метод k-средних	EM-алгоритм	Метод дерева групп
Методология	Разделение на основе схожести по евклидовому расстоянию	Разделение и распределение на основе схожести	Иерархическая группировка и объединение «снизу-вверх»
Виды входных типов данных	Реальные или случайно сгенерированные	Реальные или случайно сгенерированные	Реальные или случайно сгенерированные
Преимущества	Гибкость, эффективность и высокая скорость работы	Устойчивость к шуму в данных, возможность задать фиксированное количество групп	Применимо к любым признакам, гибкость, легкость работы с многомерными данными
Ограничения	Сложность работы с неопределенными формами иерархий.	Сложность работы с крупными наборами данных	Чувствительность к шуму, высокие требования к объему памяти и вычислений
Воспроизводимость	Различный результат	Различный результат	Одинаковый результат
Сложность	Линейная	Линейная	Квадратичная

В целом, основными минусами иерархических алгоритмов можно назвать их вычислительную сложность и требования к памяти, которые выражаются не менее чем квадратичными зависимостями. По сравнению с методом k-средних,

эти минусы являются значительными и ведут к непрактичности их применения, особенно в условиях работы с большими наборами данных [5].

6. Реализация алгоритмов

Методы многомерной классификации, которые предназначены разделять рассматриваемые совокупности объектов, субъектов или явлений на группы в определенном смысле однородные. Необходимо учитывать, что каждый из рассматриваемых объектов характеризуется большим количеством разных и стохастически связанных признаков. Для решения столь сложных задач классификации применяют кластерный и дискриминантный анализ. Наличие множества исходных признаков, характеризующих процесс функционирования объектов, заставляет отбирать из них наиболее существенные и изучать меньший набор показателей. Чаще исходные признаки подвергаются некоторому преобразованию, которое обеспечивает минимальную потерю информации. Такое решение может быть обеспечено методами снижения размерности, куда относятся факторный анализ. Этот метод позволяет учитывать эффект существенной многомерности данных, дает возможность лаконичного и более простого объяснения многомерных структур. Вскрывает объективно существующие, непосредственно не наблюдаемые закономерности при помощи полученных факторов или главных компонент.

Это дает возможность достаточно просто и точно описать наблюдаемые исходные данные, структуру и характер взаимосвязей между ними. Сжатие информации получается за счет того, что число факторов или главных компонент – новых единиц измерения – используется значительно меньше, чем исходных признаков.

Все перечисленные методы наиболее эффективны при активном применении статистических пакетов прикладных программ. При помощи этих пакетов предоставляется возможным даже восстанавливать пропущенные данные и др.

Стандартные статистические методы обработки данных включены в состав электронных таблиц, таких как Excel, Lotus 1-2-3, QuattroPro, и в математические пакеты общего назначения, например Mathcad. Но гораздо большими возможностями обладают специализированные статистические пакеты, позволяющие применять самые современные методы математической статистики для обработки данных. По официальным данным Международного статистического института, число статистических программных продуктов приближается к тысяче. Среди них есть профессиональные статистические пакеты, предназначенные для пользователей, хорошо знакомых с методами математической статистики, и есть пакеты, с которыми могут работать специалисты, не имеющие глубокой математической подготовки; есть пакеты отечественные и созданные зарубежными программистами; различаются программные продукты и по цене.

Среди программных средств данного типа можно выделить узкоспециализированные пакеты, в первую очередь статистические - STATISTICA, SPSS, STADIA, STATGRAPHICS, которые имеют большой набор статистических функций: факторный анализ, регрессионный анализ, кластерный анализ, многомерный анализ, критерии согласия и т. д. Данные программные продукты обычно содержат и средства для визуальной интерпретации полученных результатов: различные графики, диаграммы, представление данных на географической карте.

При анализе данных пользователю статистического программного пакета приходится выполнять вычисления широкого спектра статистик, передавать и преобразовывать данные для их анализа, а также представлять полученные результаты в наглядном виде [1]. Поэтому при выборе того или иного статистического пакета, для сравнения пакетов, необходимо прежде всего обращать внимание на такие характеристики, как:

- удобство управления данными (экспорт/импорт данных, их реструктуризация);

- статистическое разнообразие (количество статистических модулей);
- графические возможности (наличие встроенного графического редактора, возможность показа отдельных элементов графика, возможности экспорта графиков).

Кроме того, большое значение имеет удобство работы с пакетом, легкость его освоения (наличие встроенной системы помощи, руководства пользователя, степень удобства управления данными, результатами вычислений, таблицами и графиками), а также скорость произведения вычислений.

Существуют также нестатистические пакеты, решающие задачи классификации (PolyAnalyst, ДА-система, АРГОНАВТ, ЛОРЕГ, пакет ОТЭКС и разнообразные нейросетевые пакеты).

Для реализации рассмотренных методов многомерной классификации выбрана достаточно популярная в нашей стране система STATISTICA.

STATISTICA — программный пакет для статистического анализа, разработанный компанией Dell, реализующий функции анализа данных, управления данными, добычи данных, визуализации данных с привлечением статистических методов. Пакет обладает широкими графическими возможностями, позволяет выводить информацию в виде различных типов графиков (включая научные, деловые, трёхмерные и двухмерные графики в различных системах координат, специализированные статистические графики — гистограммы, матричные, категорированные графики и др.), все компоненты графиков настраиваются.

7. Обработка исходных данных в STATISTICA

STATISTICA обладает удобным интерфейсом для работы с крупными массивами данных. Образец представлен на рис.7.1.

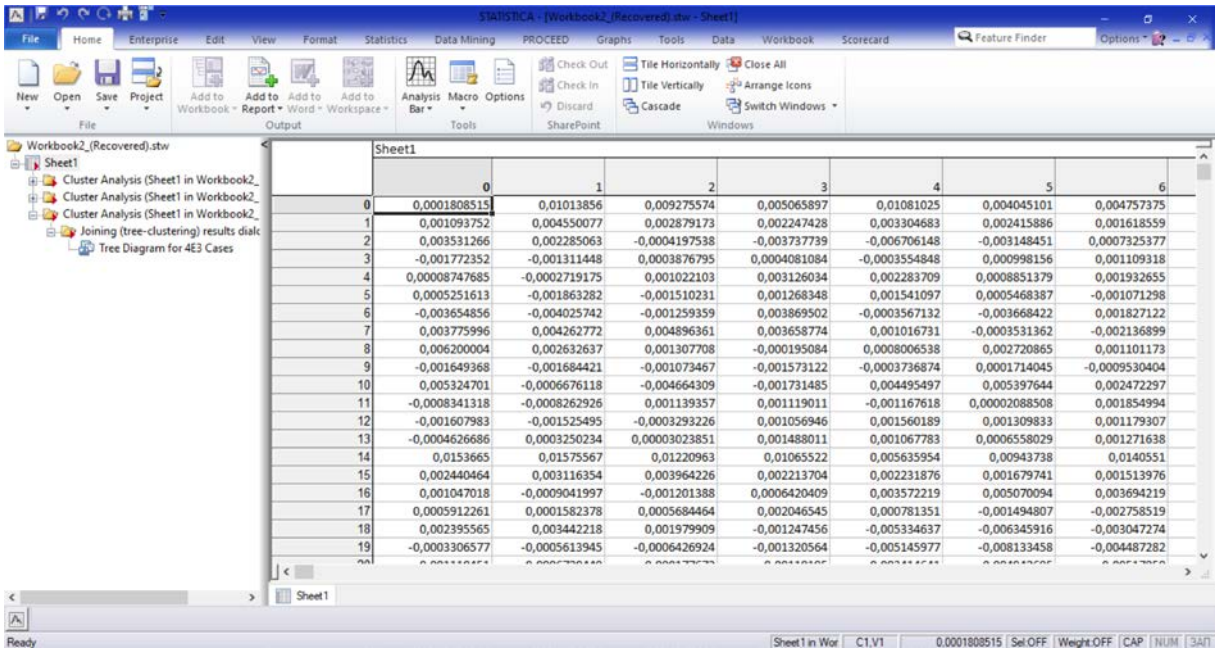

Рис.7.1 Интерфейс программы с загруженным набором данных

7.1. Нормализация данных

Первостепенной задачей является нормализация данных. В интерфейсе приложения это реализуется через пункт Standardize Variables на вкладке Edit, который представлен на рис.7.2.

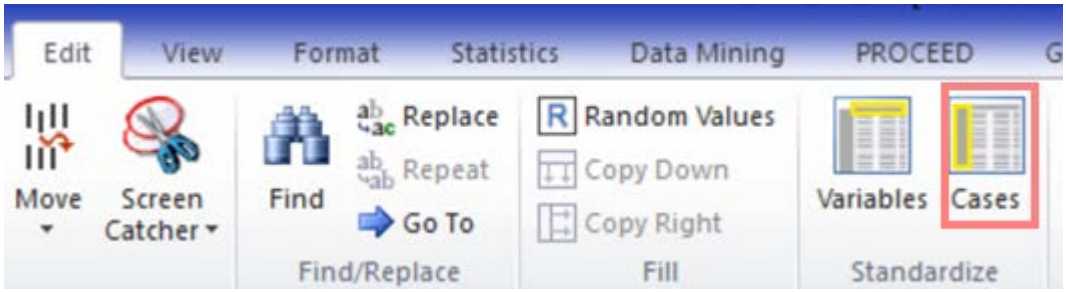

Рис.7.2 Нормализация в пакете STATISTICA

7.2. Среднеквадратическое отклонение

Следующим шагом является образование нового признака среднеквадратического отклонения. Новый столбец добавляется при помощи использования функции Variables – Add на вкладке Edit, где затем после нажатия

кнопки Funtctions появляется диалоговое окно, в котором выбирается пункт Statistics и подпункт stdev (Standard Deviation). Интерфейс представлен на рис.7.3.

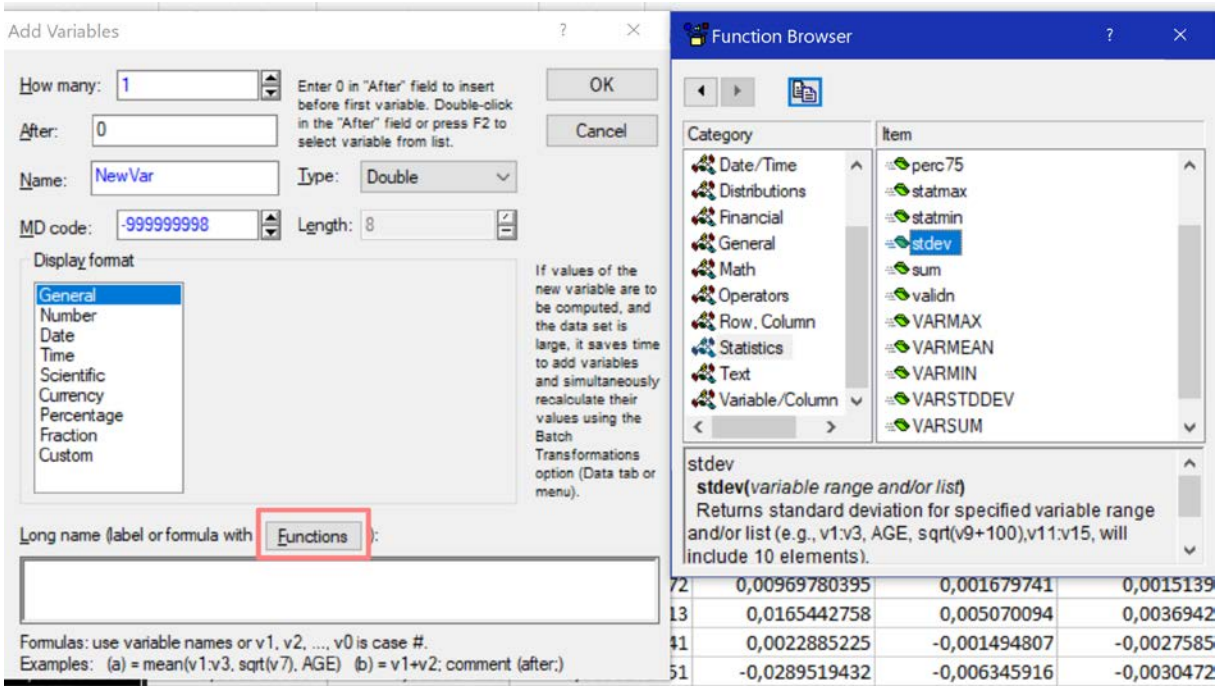

Рис.7.3 Диалоговое окно создания новых признаков

7.3. Преобразование Фурье

Затем необходимо было получить преобразование Фурье от исходных данных. В пакете прикладных программ STATISTICA это выполняется на вкладке Statistics в модуле Advanced Models в пункте Time Series/Forecasting, показано на рис.7.4.

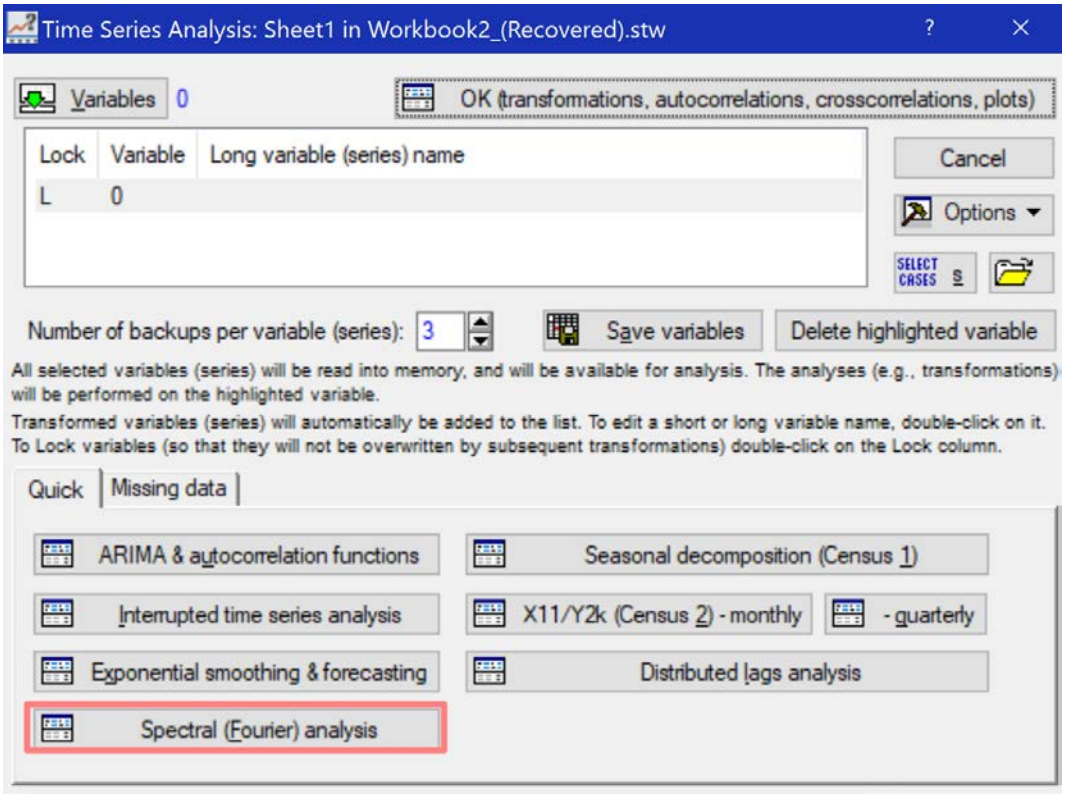

Рис.7.4 Преобразование Фурье в пакете STATISTICA

Затем к полученным величинам так же применяется взятие среднего и среднеквадратического отклонения.

8. Результаты

8.1. Метод k-средних

Первым примененным методом являлся метод k-средних. Настройка параметров метода представлена на рис.8.1.

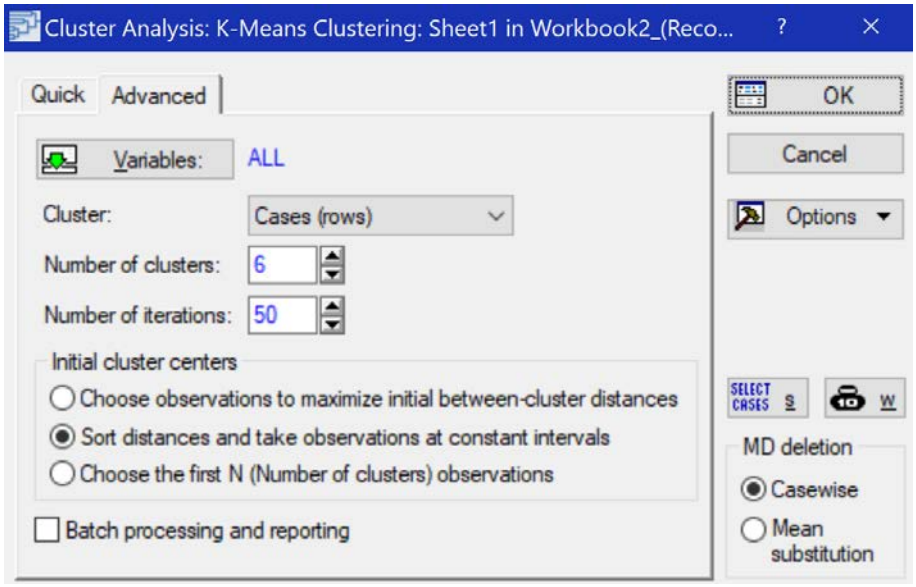

Рис.8.1 Диалоговое окно настройки метода k-средних

Результаты метода кластеризации были представлены в виде таблицы расстояний между центроидами кластеров на рис.8.2.

Cluster Number	Euclidean Distances between Clusters (Sheet1 in Workbook2_(Recovered).stw) Distances below diagonal Squared distances above diagonal					
	No. 1	No. 2	No. 3	No. 4	No. 5	No. 6
No. 1	0,000000	0,014115	0,048257	0,028879	0,045989	0,037088
No. 2	0,118807	0,000000	0,094943	0,027283	0,064583	0,054095
No. 3	0,219675	0,308128	0,000000	0,143349	0,035629	0,133272
No. 4	0,169937	0,165176	0,378614	0,000000	0,133215	0,026625
No. 5	0,214450	0,254132	0,188756	0,364986	0,000000	0,154057
No. 6	0,192582	0,232583	0,365064	0,163172	0,392501	0,000000

Рис.8.2. Результаты метода k-средних.

8.2. EM-алгоритм

Следующим использованным методом был EM-алгоритм. Его результат представлен на рис.8.3.

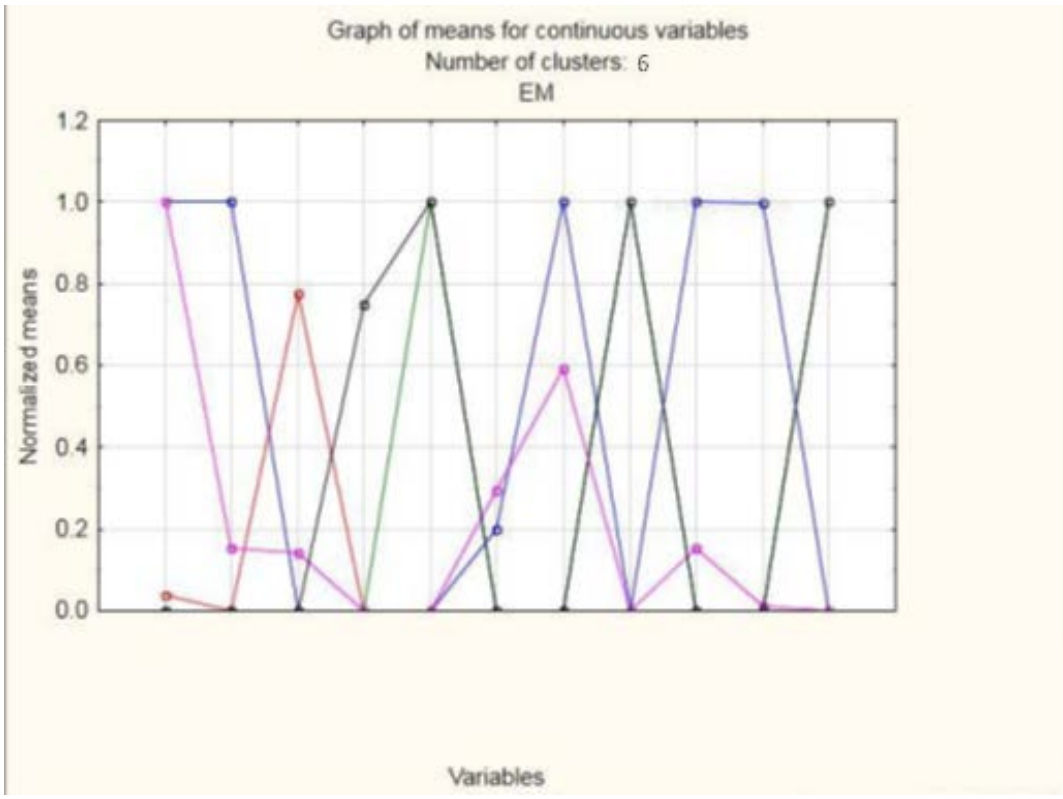

Рис.8.3 Кластеризация по EM-алгоритму

Результат работы EM-алгоритма должен быть близок к результату работы метода k-средних. EM-алгоритм показывает вероятность нахождения каждого отдельно взятого отрезка в каком-либо классе.

8.3. Метод дерева групп

Последним рассмотрен метод дерева групп. Настройки метода в программе показаны на рис. 8.4.

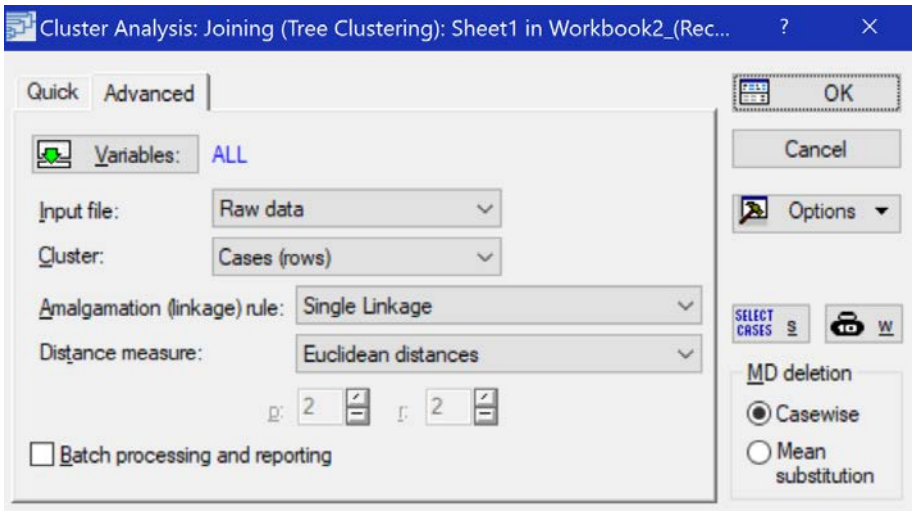

Рис.8.4 Диалоговое окно настройки метода дерева групп

Типичным представлением результатов такого метода является дерево последовательно образованных групп. Оно представлено на рис. 8.5.

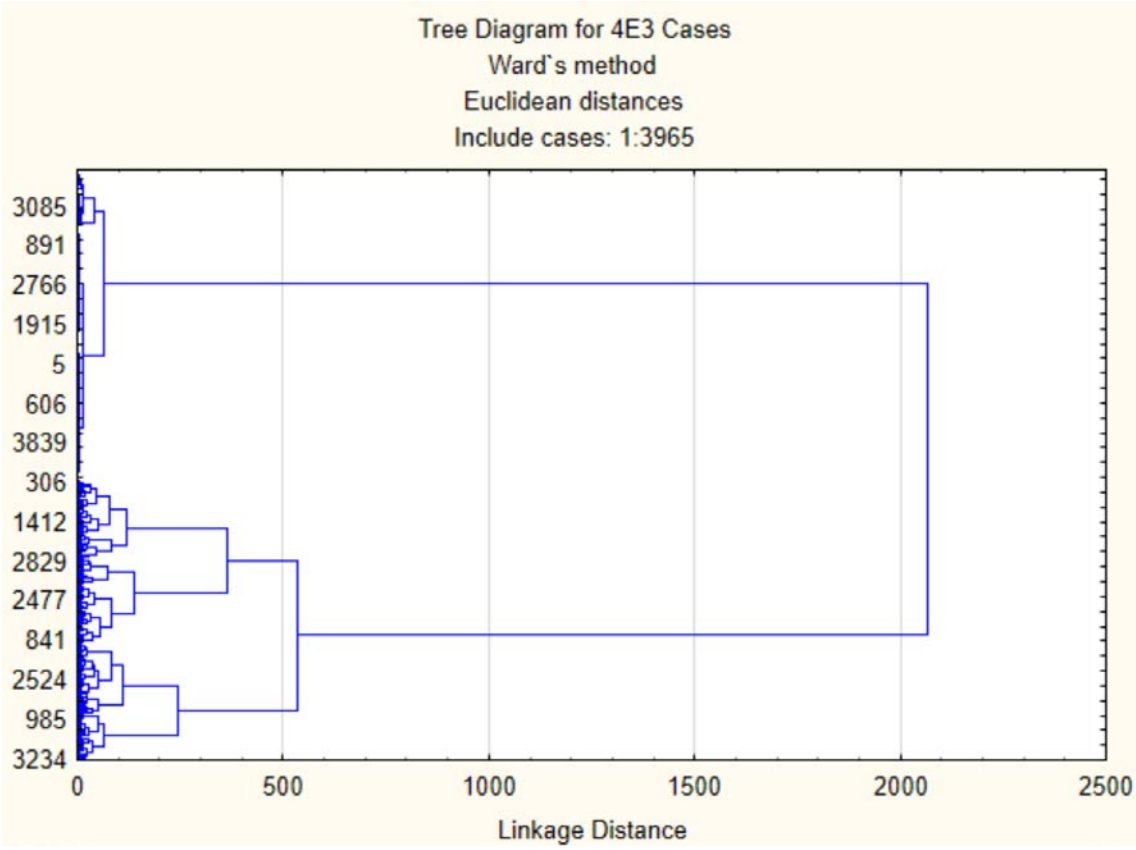

Рис.8.5 Диаграмма дерева кластеров

Метод дерева групп дал результаты отличные от двух предыдущих методов. В данном результате видно отсутствие заранее фиксированного количества кластеров. Также в конечном результате наблюдается большое различие между двумя крупными кластерами на более поздних итерациях, что может быть интерпретировано как различие между стационарными классами и классами с движением.

ЗАКЛЮЧЕНИЕ

В рамках данной выпускной работы было проведено сравнение трех методов кластерного анализа данных о движении человека.

На первом этапе были рассмотрены преимущества использования смартфона как датчика для сбора исходных данных. На следующем этапе был проведен анализ процесса извлечения признаков из исходных данных и их нормализации. Было определено минимально достаточное для работы предполагаемых методов количество признаков. Далее следовал этап выбора применяемых методов. На текущем этапе реализации задачи были выбраны три наиболее простейших метода. Для сравнения результатов работы методов был выбран пакет прикладных программ STATISTICA.

Результатом работы методов является деление обработанных отрезков исходных данных, представленных в различных видах графически и в виде таблиц. По результатам работы можно сказать, что метод k-средних справился с задачей с наибольшей точностью, что может быть объяснено использованием наперед заданного количества классов. В то же время, метод дерева групп является наиболее легко интерпретируемым, с возможностью определять активности разных масштабов сложности. Исходя из этих выводов, предлагается совместное использование метода k-средних с методом дерева групп для наиболее оптимальной работы, в рамках рассмотренных методов. Однако для такого объединения возможными проблемами могут стать вычислительная сложность и значительный объем занимаемой памяти в процессе вычисления, что характерно для попарного характера вычислений метода k-средних. Это может ограничить применимость текущих алгоритмов к наборам данным значительно большим, чем были рассмотрены в данной работе (7000 и больше отдельных промежутков обработанных данных).

В дальнейшем в рамках задачи анализа данных о движении человека предлагается рассмотреть другие алгоритмы машинного обучения, не вошедшие в данную работу. Наиболее перспективными представляются те алгоритмы,

которые естественным образом учитывают темпоральный характер задачи: DTW-алгоритм, рекуррентная нейронная сеть (LSTM-сеть).

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Буреева Н. Н. Многомерный статистический анализ с использованием ППП «STATISTICA» //Нижний Новгород. – 2007. – Т. 112.

2. Chapman A. E. Biomechanical analysis of fundamental human movements. – Human Kinetics, 2008.

3. Cui Y., Chipchase J., Ichikawa F. A cross culture study on phone carrying and physical personalization //International Conference on Usability and Internationalization. – Springer, Berlin, Heidelberg, 2007. – С. 483-492.

4. Jean-Louis G. et al. Sleep detection with an accelerometer actigraph: comparisons with polysomnography //Physiology & behavior. – 2001. – Т. 72. – №. 1-2. – С. 21-28.

5. Kolhe S., Sawarkar S. Review of document clustering techniques: issues, challenges and feasible avenue //Int J Adv Res Comput Sci Softw Eng. – 2015. – Т. 5. – №. 4.

6. Lau S. L. et al. Supporting patient monitoring using activity recognition with a smartphone //2010 7th International symposium on wireless communication systems. – IEEE, 2010. – С. 810-814.

7. Lockhart J. W., Pulickal T., Weiss G. M. Applications of mobile activity recognition //Proceedings of the 2012 ACM Conference on Ubiquitous Computing. – 2012. – С. 1054-1058.

8. Loseu V. et al. Applications of sensing platforms with wearable computers //Proceedings of the 3rd International Conference on PErvasive Technologies Related to Assistive Environments. – 2010. – С. 53:1-53:5.

9. Rajeswari K. et al. Improvement in K-means clustering algorithm using data clustering //2015 International Conference on Computing Communication Control and Automation. – IEEE, 2015. – С. 367-369.

10. Ramos-Garcia R. I., Hoover A. W. A study of temporal action sequencing during consumption of a meal //Proceedings of the International Conference on Bioinformatics, Computational Biology and Biomedical Informatics. – 2013. – С. 68:68-68:75.

11. Rao N. K., Reddy G. S. Discovery of Preliminary Centroids Using Improved K-Means Clustering Algorithm //International Journal of Computer Science and Information Technologies. – 2012. – Т. 3. – №. 3.

12. Saranya S., Zoraida B. S. E. A study on Clustering Algorithms for XML Data Clustering.

13. Scholl P. M., Van Laerhoven K. A feasibility study of wrist-worn accelerometer based detection of smoking habits //2012 Sixth International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing. – IEEE, 2012. – С. 886-891.

14. Steinbach M., Ertöz L., Kumar V. The challenges of clustering high dimensional data //New directions in statistical physics. – Springer, Berlin, Heidelberg, 2004. – С. 273-309.

15. Villar J. R. et al. Improving human activity recognition and its application in early stroke diagnosis //International journal of neural systems. – 2015. – Т. 25. – №. 04. – С. 1-20.

16. Watkins J. Structure and function of the musculoskeletal system. Champaign, Il //J Hum Kinet. – 1999. – Т. 131. – С. 181.

17. Yan Z. et al. Energy-efficient continuous activity recognition on mobile phones: An activity-adaptive approach //2012 16th international symposium on wearable computers. – Ieee, 2012. – С. 17-24.