Федеральное государственное бюджетное образовательное учреждение
высшего образования
«Новосибирский государственный технический университет»

На правах рукописи

Уваров Вадим Евгеньевич

РАЗРАБОТКА И ИССЛЕДОВАНИЕ МЕТОДОВ РАСПОЗНАВАНИЯ ПОСЛЕДОВАТЕЛЬНОСТЕЙ, ОПИСЫВАЕМЫХ СКРЫТЫМИ МАРКОВСКИМИ МОДЕЛЯМИ, ПРИ НЕПОЛНЫХ ДАННЫХ

Специальность: 05.13.17 – Теоретические основы информатики

Диссертация на соискание ученой степени
кандидата технических наук

Научный руководитель:
доктор технических наук, профессор
Попов Александр Александрович

Новосибирск – 2019

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ

Актуальность темы исследования. Тема диссертационной работы является актуальной, поскольку во многих прикладных задачах, связанных с обработкой информации, возникает потребность в анализе потоков данных от различных датчиков в сложной помеховой обстановке, когда возможно пропадание информации или ее искажение. Такие условия наблюдаются при распознавании звуков или речи при сильных посторонних шумах, при анализе биологических последовательностей, имеющих малую надёжность, например, цепочек ДНК, а также в сложных системах, например, при приеме данных с космических и летательных аппаратов и других источников.

В качестве надёжного инструмента анализа потоков информации, формализованных в виде символьных или многомерных числовых последовательностей, хорошо себя зарекомендовали скрытые марковские модели (далее – СММ). Тем не менее, в теории СММ имеется практически неизученная область, которая касается способов применения СММ в случае неполных данных, когда значение некоторых наблюдений в последовательности не определено, т. е. имеются пропуски, причем предполагается, что пропуски возникают в случайных местах последовательности без какой-либо закономерности. Отсутствие универсальных, подтверждённых теоретически и экспериментально методов использования СММ в ситуациях информационной неопределённости препятствует эффективному использованию СММ для решения задач, предполагающих наличие ненадёжных или зашумлённых данных, что определяет необходимость разработки методов анализа неполных последовательностей, описываемых СММ.

Степень разработанности **темы исследования.** Концепция скрытых марковских моделей (СММ) была предложена ещё в 1970-х годах коллективом учёных во главе с Л. Баумом [1-5]. Традиционно СММ применялись для распознавания речи [6-10]. Начиная с 1980-х годов СММ стали применять в биоинформатике [11-14], например, при анализе цепочек ДНК [15]. Также СММ успешно применялись для моделирования экономических процессов [16-19] и в задачах компьютерного зрения [20-23]. Наибольшей популярностью СММ стали пользоваться после 1990-х

годов [24], и данная тенденция сохранилась вплоть до настоящего времени, что можно подтвердить частотой упоминания термина “hidden Markov model” в публикациях [25]. Одним из недавних способов применения СММ для моделирования, являются задачи распознавания двигательной активности человека. Этот класс задач включает в себя как распознавание совершаемого движения [26-28], так и идентификацию субъекта, совершающего движение [29-31]. Также СММ хорошо зарекомендовали себя при решении задач декодирования оптимального маршрута по последовательности геоданных [32-36].

Проблема использования СММ для анализа неполных последовательностей частично освещается в статье [37], где с помощью СММ решалась задача распознавания зашумлённой речи. В цитируемой работе анализировались спектрограммы, которые были получены с помощью оконного преобразования Фурье на основе записей речи, содержащих помехи. Авторы предложили в дополнение к классическим методам фильтрации шума, использовать метод, который основан на том, что отдельные сильнозашумленные участки спектрограммы считаются утерянными. Распознавание подобных последовательностей проводилось с использованием двух методов: маргинализации пропущенных наблюдений и предварительного восстановления последовательностей. Авторы показали, что подобные методы показывают лучший результат при распознавании зашумлённой речи, чем классические методы фильтрации шумов. Результаты другого исследования, в котором проводилось распознавание неполных последовательностей с помощью СММ представлены в [38]. В данной работе рассматривалась задача распознавания движений человека по видеоряду и их воспроизведения виртуальной моделью, изображающей человека. Пропуск наблюдений в этом случае обуславливался тем, что часть тела человека, движения которого повторяет модель, могла быть невидима, – к примеру, закрыта препятствием. Для распознавания неполных последовательностей также задействовался метод маргинализации пропусков, а для определения последовательности движений человека использовался алгоритм декодирования неполных последовательностей.

Тем не менее, в упомянутых выше работах тема анализа неполных последовательностей, описываемых СММ, затронута лишь частично. Авторы не освещают вопросы обучения СММ по неполным последовательностям, теоретически не обосновывают используемые методы и не проводят сравнительный анализ их эффективности, преимуществ и недостатков. К тому же предлагаемые ими методы ограничены исключительно конкретной предметной областью: распознаванием речи и распознаванием движений по видеоряду. Поэтому данная тема нуждается в дальнейшей разработке.

Объектом исследования диссертационной работы являются методы анализа последовательностей, описываемых скрытыми марковскими моделями.

Предметом исследования диссертационной работы являются методы анализа неполных последовательностей, описываемых скрытыми марковскими моделями.

Цель и задачи исследования. Основной целью диссертационной работы является разработка и исследование методов анализа неполных последовательностей, описываемых скрытыми марковскими моделями.

Для достижения поставленной цели предусмотрено решение следующих задач. Разработать и исследовать методы:

– восстановления и декодирования неполных последовательностей, описываемых скрытыми марковскими моделями;

– распознавания неполных последовательностей, описываемых скрытыми марковскими моделями;

– обучения скрытой марковской модели по неполным последовательностям;

– распознавания неполных последовательностей, описываемых близкими скрытыми марковскими моделями, обученными на неполных последовательностях.

Идея диссертационной работы заключается в использовании маргинального распределения непропущенных наблюдений путем интегрирования совместного распределения пропущенных и непропущенных наблюдений по всем возможным

значениям пропущенных наблюдений для анализа неполных последовательностей, описываемых скрытыми марковскими моделями.

Научная новизна диссертационной работы заключается в том, что **впервые разработаны и исследованы**:

– метод восстановления и декодирования неполных последовательностей, описываемых скрытыми марковскими моделями, основанный на модифицированном алгоритме Витерби;

– метод распознавания неполных последовательностей, описываемых скрытыми марковскими моделями, основанный на модифицированном алгоритме forward-backward;

– метод обучения скрытой марковской модели по неполным последовательностям, основанный на модифицированном алгоритме Баума-Велша;

– метод распознавания неполных последовательностей, основанный на модифицированном алгоритме вычисления производных от логарифма функции правдоподобия того, что случайный процесс, описываемый скрытой марковской моделью, сгенерировал подобную последовательность.

Личный вклад автора заключается в том, что **автором лично**:

– разработаны методы на основе: модифицированный алгоритм Витерби, модифицированный алгоритм forward-backward, модифицированный алгоритм Баума-Велша и модифицированный алгоритм вычисления первых производных от логарифма правдоподобия того, что случайный процесс, описываемый скрытой марковской моделью, сгенерировал неполную последовательность;

– проведены вычислительные эксперименты, анализ их результатов и сделаны выводы;

– реализованы описанные в диссертационной работе алгоритмы, а также вычислительные эксперименты в программе для ЭВМ;

– разработаны практические методики:

1) декодирования наиболее вероятного пути движения абонента по транспортному графу на основе последовательности регистраций в мобильной сети, используемая в работе оператора сотовой связи Tele2 компанией ООО «Т2 Мобайл»;

2) восстановления неполных данных двигательной активности человека;

3) идентификации личности по неполным данным двигательной активности при полной и неполной обучающих выборках, а также с использованием производных;

– оценена эффективность разработанных методов и даны рекомендации по проведению дальнейших исследований.

Теоретическая значимость. Исследования, проведённые в диссертации, позволяют расширить раздел теоретической информатики, касающийся анализа последовательностей, описываемых скрытыми марковскими моделями, применительно к случаю наличия пропусков в последовательностях.

Практическая значимость. Программа для ЭВМ, предложенная автором на основе разработанных алгоритмов позволяет решать практические задачи анализа неполных последовательностей, порождённых случайными процессами, описываемыми скрытыми марковскими моделями, таких как данные с акселерометра носимого устройства, а также последовательности координат с GPS-устройства, либо устройств мобильной связи;

Методология и методы исследования. Теоретической базой исследования являются методы теории машинного обучения, теории вероятностей, математической статистики и математического анализа. Для решения поставленных задач использовалось статистическое моделирование, экспериментальные исследования, а также сравнительный анализ эффективности алгоритмов.

Положения, выносимые на защиту:

– метод на основе модифицированного алгоритма Витерби позволяет проводить декодирование неполных последовательностей, описываемых скрытыми марковскими моделями до 1.4 раза, а восстановление в них пропусков до 7 раз точнее, чем при использовании альтернативных методов СММ.

– метод на основе модифицированного алгоритма forward-backward позволяет проводить распознавание неполных последовательностей, описываемых скрытыми марковскими моделями, до 1.6 раз точнее, чем при использовании стандартных методов СММ.

– метод на основе модифицированного алгоритма Баума-Велша позволяет проводить обучение скрытых марковских моделей по неполным последовательностям до 1.2 раз эффективнее других известных методов СММ.

– метод на основе модифицированного алгоритма вычисления первых производных от логарифма функции правдоподобия того, что случайный процесс, описываемый скрытой марковской моделью, сгенерировал неполную последовательность, позволяет до 1.2 раз повысить количество правильно классифицированных неполных последовательностей по сравнению с другими методами СММ.

Обоснованность и достоверность научных положений, выводов и рекомендаций обеспечивается:

– базированием на строго доказанных и корректно используемых постулатах теоретической информатики, что подтверждает непротиворечивость разработанных автором теоретических моделей уже известным научным положениям;

– корректным применением методов машинного обучения, теории вероятностей, математической статистики и математического анализа;

– подтверждением эффективности разработанных методов представительной выборкой результатов вычислительных экспериментов и положительным их применением для решения практических задач.

Апробация работы. Основные результаты исследований, проведенных автором, докладывались и обсуждались: на XIII международной научно-технической конференции “Актуальные проблемы электронного приборостроения” АПЭП-2016 (International Scientific and Technical Conference on Actual Problems of Electronic Instrument Engineering APEIE-2016) 3-6 октября 2016 года в г. Новосибирск; на международной конференции “Прикладные методы статистического анализа: непараметрические подходы в кибернетике и системном анализе” (International Workshop on Applied Methods of Statistical Analysis: Nonparametric Methods in Cybernetics and System Analysis AMSA-2017) в г. Красноярск 17-22 сентября 2017 года; на XI Международной IEEE научно-технической конференции “Динамика систем, механизмов и машин” Динамика-2017 (International Scientific and Technical Conference on

Applied Mechanics and Dynamics Systems Dynamics-2017) в г. Омск 14-16 ноября 2017 года; на XII Международной IEEE научно-технической конференции “Динамика систем, механизмов и машин” Динамика-2018 (International Scientific and Technical Conference on Applied Mechanics and Dynamics Systems Dynamics-2018) в г. Омск 13-15 ноября 2018 года; на международной конференции “Прикладные методы статистического анализа: непараметрический подход” (International Workshop on Applied Methods of Statistical Analysis: Nonparametric Approach AMSA-2015) в г. Белокуриха 14-19 сентября 2015 года; на городской научно-практической конференции аспирантов и магистрантов “Progress Through Innovations” в г. Новосибирск 31 марта 2016 года; на городской научно-практической конференции студентов, магистрантов и аспирантов “Aspire to Science” в г. Новосибирск 12 марта 2016 года; на российской научно-технической конференции «Обработка информации и математическое моделирование» ОИиММ-2016 в г. Новосибирск 21-22 апр. 2016; на 11-м международном форуме по стратегическим технологиям IFOST-2016 (11-th International Forum on Strategic Technology IFOST-2016) в г. Новосибирск 1-3 июня 2016 года; на всероссийской научной конференции молодых ученых «Наука. Технологии. Инновации» НТИ-2016 в г. Новосибирск 5-9 декабря 2016; на российской научно-технической конференции «Обработка информации и математическое моделирование» ОИиММ-2017 в г. Новосибирск 26-27 апр. 2017 года.

Реализация полученных результатов.

Результаты диссертационных исследований использованы при внедрении системы отслеживания передвижения абонентов, включающей разработанный автором новый метод для привязки (map-matching) треков передвижения пользователей устройств мобильной связи к транспортному графу. Метод разработан и эффективно применяется на предприятии ООО «Т2 Мобайл», г. Москва, что подтверждено соответствующей актом об использовании результатов диссертационной работы.

Публикации. Основные научные результаты диссертации опубликованы в 16 печатных работах, из которых 4 – в изданиях, входящих в «Перечень ведущих

рецензируемых научных журналов и изданий, в которых должны быть опубликованы основные научные результаты диссертаций на соискание учёной степени доктора и кандидата наук», 5 – в изданиях, индексируемых в базах данных Web Of Science и Scopus, 7 – в сборниках научных работ и материалах конференций, индексируемых РИНЦ. Имеется одно свидетельство о государственной регистрации программы для ЭВМ.

Структура работы. Диссертация состоит из введения, 5 глав, заключения, списка сокращений, списка условных обозначений, словаря терминов, списка литературы (100 источников) и 2 приложений. Основной текст работы изложен на 134 страницах, включает 2 таблицы и 28 рисунков.

Краткое содержание работы. В первой главе дано современное состояние проблемы анализа неполных последовательностей, описываемых скрытыми марковскими моделями, приводится структура скрытой марковской модели, рассмотрены известные алгоритмы распознавания и декодирования последовательностей, описываемых СММ, а также обучения СММ. Освещены вопросы моделирования неполных последовательностей, а также вычисления производных от функции правдоподобия того, что описываемый СММ процесс сгенерировал последовательность. Первая глава завершается постановкой задач исследования.

В последующих главах даны разработанные автором методы и практические методики анализа неполных последовательностей, описываемых СММ.

Во второй главе описан разработанный автором метод на основе модифицированного алгоритма Витерби, позволяющий осуществлять декодирование неполных последовательностей, описываемых скрытыми марковскими моделями, научно обосновано применение данного алгоритма для восстановления неполных последовательностей, описываемых скрытыми марковскими моделями, а также проведен сравнительный анализ эффективности разработанного алгоритма и других методов декодирования и восстановления неполных последовательностей, описываемых скрытыми марковскими моделями. Также в этой главе приведены две новые практические методики: восстановления неполных данных двигательной активности человека, а также методика декодирования наиболее вероятного пути

движения абонента по транспортному графу на основе последовательности регистраций в мобильной сети.

В третьей главе дан новый метод на основе модифицированного алгоритма forward-backward, который позволяет вычислять значение логарифма функции правдоподобия того, что случайный процесс, описываемый СММ сгенерировал неполную последовательность, обосновано применение метода для распознавания неполных последовательностей, описываемых скрытой марковской моделью, а также проводится сравнительный анализ эффективности разработанного алгоритма и других методов распознавания неполных последовательностей, описываемых скрытыми марковскими моделями, в том числе и описанного во второй главе. Кроме того, в этой главе приведена новая практическая методика идентификации личности по неполным данным двигательной активности при полной обучающей выборке.

В четвёртой главе описан разработанный автором метод на основе модифицированного алгоритма Баума-Велша, позволяющий производить обучение скрытых марковских моделей по неполным последовательностям, а также результаты сравнительного анализа эффективности разработанного метода и других методов обучения СММ по неполным последовательностям. Также в этой главе дана новая практическая методика идентификации личности по неполным данным двигательной активности при неполной обучающей выборке.

В пятой главе автором предложен метод распознавания неполных последовательностей, описываемых скрытыми марковскими моделями, обученными на неполных последовательностях на основе модифицированного алгоритма вычисления первых производных от логарифма функции правдоподобия того, что случайный процесс, описываемый скрытой марковской моделью, сгенерировал неполную последовательность. Проведен сравнительный анализ эффективности разработанного метода и других методов распознавания неполных последовательностей, описываемых СММ. Кроме того, в этой главе приведена новая практическая методика идентификации личности по неполным данным двигательной активности с использованием производных от логарифма функции правдоподобия.

ГЛАВА 1 ИССЛЕДОВАНИЕ СОВРЕМЕННОГО СОСТОЯНИЯ ПРОБЛЕМЫ

В данной главе изучено современное состояние проблемы анализа неполных последовательностей, описываемых скрытыми марковскими моделями, приведена структура скрытой марковской модели, рассмотрены известные алгоритмы распознавания и декодирования последовательностей, описываемых скрытыми марковскими моделями, а также обучения скрытых марковских моделей. Освещены вопросы моделирования неполных последовательностей, а также вычисления производных от функции правдоподобия того, что описываемый скрытой марковской моделью процесс сгенерировал последовательность. Первая глава завершается постановкой задач исследования.

1.1 Обзор источников по теме диссертационной работы

Концепция скрытых марковских моделей (СММ) была предложена ещё в 1970-х годах коллективом учёных во главе с Л. Баумом [1-5]. Традиционно СММ применялись для распознавания речи [6-10]. Начиная с 1980-х годов СММ стали применять в биоинформатике [11-14], например, при анализе цепочек ДНК [15]. Также СММ успешно применялись для моделирования экономических процессов [16-19] и в задачах компьютерного зрения [20-23]. Наибольшей популярностью СММ стали пользоваться после 1990-х годов [24], и данная тенденция сохранилась вплоть до настоящего времени, что можно подтвердить частотой упоминания термина “hidden Markov model” в публикациях [25]. Одним из недавних способов применения СММ для моделирования, являются задачи распознавания двигательной активности человека. Этот класс задач включает в себя как распознавание совершаемого движения [26-28], так и идентификацию субъекта, совершающего движение [29-31]. Также СММ хорошо зарекомендовали себя при решении задач декодирования оптимального маршрута по последовательности геоданных [32-36].

Во многом популярность концепции скрытых марковских моделей возросла благодаря развитию вычислительных технологий и повышению скорости обра-

ботки данных. Помимо непосредственного увеличения тактовой частоты процессоров, появились различные инструменты распределения вычислительной нагрузки, такие как мультиядерные процессоры, мультикомпьютеры, а также различные ускорители: видеокарты и сопроцессоры. Обучение СММ на объемных выборках порой может занимать недели, поэтому актуальное направление исследований заключается в разработке параллельных реализаций алгоритмов работы с СММ на различных вычислительных устройствах. Параллельная реализация алгоритмов работы с СММ на многоядерном процессоре была представлена, например, в работе [39], где авторы получили ускорение в 5.5 раз по сравнению с последовательной версией. В работе [40] алгоритмы СММ были реализованы с помощью видеокарты и было достигнуто ускорение от 4 до 17 раз по сравнению с последовательной реализацией. В другой работе [41] авторы смогли достичь с помощью видеокарты ускорения алгоритма расчета апостериорной вероятности последовательности в 800 раз, а алгоритма Баума-Велша в 200 раз по сравнению с последовательной версией, реализованной на центральном процессоре. Также на видеокарте был реализован и алгоритм Витерби [42]. Автор данной реализации алгоритмов работы с СММ достиг ускорения в 300 раз по сравнению с последовательной версией. Также исследования в области ускорения алгоритмов СММ с помощью видеокарт велись автором данной диссертационной работы и его соавторами [43-50].

В последнее время широко исследуется такие направления в теории СММ как инкрементальное и онлайн обучение скрытых марковских моделей. Инкрементальный подход предполагает обработку обучающей последовательности по частям, что позволяет, во-первых, экономить память, необходимую для проведения вычислений, а во-вторых, ускоряет сходимость оценок параметров к истинным параметрам. Онлайн подход в свою очередь, предполагает, что обучающая информация доступна не сразу, а становится доступна поблочно или посимвольно, из чего следует, что СММ необходимо обучать в режиме реального времени, адаптируя её параметры к новым поступившим данным. Весьма подробный обзор различных методик инкрементального и онлайн обучения, включающих в себя как поблочные, так и посимвольные, а также основанных как на алгоритме Баума-Велша, так и на

прямых алгоритмах, представлен в работе [51]. Кроме того, там отмечалась способность таких алгоритмов адаптироваться к новым данным быстрее, тем самым избегая так называемой ловушки локального максимума. Улучшенная инкрементальная версия алгоритма Баума Велша, основанная на аппроксимации обратной переменной, была разработана и представлена в работе [52]. Авторы данной работы показали, что инкрементальная версия алгоритма Баума-Велша обеспечивает сходимость оценок параметров модели к истинным гораздо быстрее, чем стандартная версия. Кроме того, инкрементальные СММ были применены к задаче распознавания числовых и буквенных символов [53]. Авторы использовали набор из СММ для проведения исследований, а обучение производили с помощью инкрементального подхода, тем самым получив лучшие результаты, чем при традиционном подходе.

Как известно, для описания наблюдений с помощью СММ как правило используются смеси нормальных распределений. Данный подход достаточно гибок и позволяет добиться хороших результатов, однако, он достаточно чувствителен к выбросам. В работе [54] было предложено использовать для описания распределения наблюдений смесь распределений Стьюдента, что позволяет повысить робастность метода. Распределение Стьюдента имеет дополнительный параметр, чем отличается от нормального распределения. При устремлении же этого параметра к бесконечности, распределение Стьюдента сходится к нормальному распределению. Таким образом, по сути, данный метод является обобщением СММ с нормальными распределениями. Другой проблемой, требующей исследования, является проблема повышения дискриминантной способности СММ в условиях близких параметров. Проблема состоит в том, что традиционный методы классификации с использованием СММ, например, критерий максимума функции правдоподобия, в таких условиях не способен предоставить качественное различие между последовательностями, принадлежащими различным СММ. Метод, предложенный Поповым А. А. и Гультяевой Т. А. (Новосибирский государственный технический университет) состоит в использовании производных от функции правдоподобия по параметрам СММ в качестве признаков для проведения анализа другими классификаторами, в частности — методом опорных векторов [55-59].

Как уже было замечено, в теории СММ имеется практически неизученная область, которая касается способов применения СММ в случае неполных данных. Неполная последовательность наблюдений – это такая последовательность, где значение некоторых наблюдений не определено, причем предполагается, что пропуски возникают в случайных местах последовательности без какой-либо закономерности. Частично данная проблема затрагивается в статье [37], где с помощью СММ решалась задача распознавания зашумлённой речи. В цитируемой работе анализировались спектрограммы, которые были получены с помощью оконного преобразования Фурье на основе записей речи, содержащих помехи. Авторы предложили в дополнении к классическим методам фильтрации шума, использовать подход, который основан на том, что отдельные сильно зашумленные участки спектрограммы считаются утерянными. Распознавание подобных последовательностей проводилось с использованием двух подходов: маргинализации пропущенных наблюдений и предварительного восстановления последовательностей. Авторы показали, что подобные подходы показывают лучший результат при распознавании зашумлённой речи, чем классические методы фильтрации шумов. Результаты другого исследования, в котором проводилось распознавание неполных последовательностей с помощью СММ представлены в [38]. В данной работе рассматривалась задача распознавания движений человека по видеоряду и их воспроизведения виртуальной моделью, изображающей человека. Пропуск наблюдений в этом случае обуславливался тем, что часть тела человека, движения которого повторяет модель, могла быть невидима, – к примеру, закрыта препятствием. Для распознавания неполных последовательностей также был задействован подход маргинализации пропусков, а для определения последовательности движений человека использовался алгоритм декодирования неполных последовательностей.

Тем не менее, в упомянутых выше работах тема анализа неполных последовательностей, описываемых СММ, затронута лишь частично. Авторы не освещают вопросы обучения СММ по неполным последовательностям, не проводят теоретическое исследование используемых подходов и не проводят сравнительный анализ

эффективности этих подходов и других подходов анализа неполных последовательностей, описываемых СММ, не выявляют преимущества и недостатки этих подходов. К тому же предлагаемые ими подходы тесно привязаны к конкретной предметной области: распознаванию речи и распознаванию движений по видеоряду.

1.2 Понятие «скрытая марковская модель»

Скрытой марковской моделью называют модель, описывающую случайный процесс, находящийся в каждый момент времени $t \in \{1, ..., T\}$ в одном из N скрытых состояний $s \in \{s_1, ..., s_N\}$ и в новый момент времени переходящий в другое или в прежнее состояние согласно некоторым вероятностям переходов. В момент нахождения в очередном состоянии процесс, описываемый СММ, генерирует видимое наблюдение. Состояния считаются скрытыми, однако они проявляются в тех или иных особенностях генерируемых последовательностей наблюдений. Настоящее исследование ограничивается рассмотрением СММ, в которых распределение очередного состояния зависит только от одного предыдущего состояния.

В данной работе рассматриваются СММ, как с дискретным распределением наблюдений, где наблюдения принадлежат конечному алфавиту символов, так и с непрерывной плотностью распределения наблюдений, когда в общем случае наблюдения – это векторы действительных чисел. Значения наблюдаемых величин при условии того, что СММ находится в конкретном скрытом состоянии, подчиняются некоторым вероятностным законам. В случае СММ с дискретным распределением наблюдений эти вероятностные законы описываются условными дискретными распределениями, а в случае СММ с непрерывной плотностью распределения наблюдений – функциями условной плотности распределений наблюдений.

Рассмотрим параметры, которыми можно полностью задать конкретную СММ. Обозначим скрытое состояние, в котором находится описываемый СММ процесс в момент t символом q_t, а текущее скрытое состояние, не привязанное к времени, символом q; одно из скрытых состояний (например, под номером i), в

котором может находиться процесс, символом s_i; наблюдение, которое он сгенерировал в момент времени t символом $\boldsymbol{o}_t$, а наблюдение, не привязанное к конкретному моменту времени, символом $\boldsymbol{o}$.

СММ с дискретным распределением наблюдений, имеющую N скрытых состояний и алфавит наблюдений, состоящий из M символов, можно представить вектором вероятностного распределения начального скрытого состояния $\Pi = \left\{\pi_i = p\left(q_1 = s_i\right),\ i = \overline{1,N}\right\}$, матрицей вероятностей переходов из одного скрытого состояния в другое $A = \left\{a_{ij} = p\left(q_{t+1} = s_j \mid q_t = s_i\right),\ \ i, j = \overline{1,N}\right\}$, а также условным дискретным распределением наблюдений: $B = \left\{b_i\left(\boldsymbol{o}\right) = p\left(\boldsymbol{o} \mid q = s_i\right), i = \overline{1,N}, \boldsymbol{o} \in V\right\}$, $V = \{v_1, v_2, \ldots, v_M\}$, где V - конечный алфавит символов, а $v_m, m = \overline{1,M}$ - символы алфавита. При этом соблюдаются следующие ограничения на значения параметров, исходя из их вероятностной природы: $\pi_i \geq 0, i = \overline{1,N}$; $\sum_{i=1}^{N} \pi_i = 1$;

$a_{ij} \geq 0,\ i, j = \overline{1,N}$; $\sum_{j=1}^{N} a_{ij} = 1,\ \ i = \overline{1,N}$; $b_i\left(v\right) \geq 0 = 1, i = \overline{1,N}, v \in V$; $\sum_{v \in V} b_i\left(v\right) = 1, i = \overline{1,N}$.

СММ с непрерывной плотностью распределения наблюдений, имеющая N скрытых состояний, характеризуется вектором вероятностного распределения начального скрытого состояния $\Pi = \left\{\pi_i = p\left(q_1 = s_i\right),\ i = \overline{1,N}\right\}$, матрицей вероятностей переходов из одного скрытого состояния в другое $A = \left\{a_{ij} = p\left(q_{t+1} = s_j \mid q_t = s_i\right),\ \ i, j = \overline{1,N}\right\}$, а также функциями условной плотности распределений многомерных наблюдений: $B = \left\{b_i\left(\boldsymbol{o}\right) = f\left(\boldsymbol{o} \mid q = s_i\right), i = \overline{1,N}, \boldsymbol{o} \in R^Z\right\}$. В данной работе в качестве функций условной плотности распределения наблюдений рассматривается смесь многомерных нормальных распределений:

$$b_i(\boldsymbol{o}) = \sum_{m=1}^{M} \tau_{im} g(\boldsymbol{o}; \mu_{im}, \Sigma_{im}),\ \ i = \overline{1,N}, \boldsymbol{o} \in R^Z,$$

где M – число компонент в смеси для каждого скрытого состояния, $\tau_{im}, i=1,N, m=\overline{1,M}$ – вес m-й компоненты смеси в i-м скрытом состоянии, μ_{im} – математическое ожидание нормального распределения, соответствующего m-й компоненте смеси в i-м скрытом состоянии, Σ_{im} – ковариационная матрица нормального распределения, соответствующая m-й компоненте смеси в i-м скрытом состоянии, а $g(\boldsymbol{o};\mu_{im},\Sigma_{im}), \boldsymbol{o}\in R^Z$ – функция плотности многомерного нормального распределения, т.е. $g(\boldsymbol{o};\mu_{im},\Sigma_{im})=\frac{1}{\sqrt{(2\pi)^Z|\Sigma_{im}|}}e^{-0.5(\boldsymbol{o}-\mu_{im})^T\Sigma_{im}^{-1}(\boldsymbol{o}-\mu_{im})}, \boldsymbol{o}\in R^Z$. При этом соблюдаются следующие ограничения на значения параметров, исходя из их вероятностной природы: $\pi_i\geq 0, i=\overline{1,N}$; $\sum_{i=1}^{N}\pi_i=1$; $a_{ij}\geq 0,\ i,j=\overline{1,N}$; $\sum_{j=1}^{N}a_{ij}=1,\ \ i=\overline{1,N}$; $\tau_{im}\geq 0, i=1,N, m=\overline{1,M}$; $\sum_{m=1}^{M}\tau_{im}=1,\ i=\overline{1,N}$; $\Sigma_{im}\geq 0,\ i=1,N, m=\overline{1,M}$.

Таким образом, некоторую конкретную СММ будем задавать в виде набора определяющих её параметров $\lambda=\{\Pi,A,B\}$, где параметр B будет иметь различный состав в зависимости от типа СММ [60].

1.3 Декодирование последовательностей, описываемых скрытыми марковскими моделями

Для декодирования последовательностей, порождённых процессами, описываемыми СММ, то есть формирования наиболее вероятной последовательности скрытых состояний $\hat{Q}=\{\hat{q}_1,\ldots,\hat{q}_T\}$ по наблюдаемой последовательности $O=\{\boldsymbol{o}_1,\ldots,\boldsymbol{o}_T\}$, традиционно используется эффективный алгоритм Витерби, который выглядит следующим образом [61-62].

Алгоритм Витерби:

1) инициализация:

$$\delta_1(i)=\pi_i b_i(\boldsymbol{o}_1),\quad i=\overline{1,N}, \tag{1}$$

$$\psi_1(i)=0,\quad i=\overline{1,N}; \tag{2}$$

2) индукция:

$$\delta_t(j) = \max_{1 \le i \le N} \left[\delta_{t-1}(i) a_{ij} \right] b_j(\boldsymbol{o}_t), \quad j = \overline{1,N}, \quad t = \overline{2,T} \tag{3}$$

$$\psi_t(j) = \arg\max_{1 \le i \le N} \left[\delta_{t-1}(i) a_{ij} \right], \quad j = \overline{1,N}, \; t = \overline{2,T} \tag{4}$$

3) завершение:

$$\hat{q}_T = \arg\max_{1 \le i \le N} \left[\delta_T(i) \right], \tag{5}$$

4) рекурсивное определение наиболее вероятной последовательности скрытых состояний:

$$\hat{q}_t = \psi_{t+1}(\hat{q}_{t+1}), \quad t = \overline{T-1, \; 1}. \tag{6}$$

После завершения алгоритма получим сформированную наиболее вероятную последовательность скрытых состояний: $\hat{Q} = \{\hat{q}_1, \cdots, \hat{q}_T\}$.

Рисунок 1 содержит иллюстрацию работы алгоритма Витерби для СММ, состоящей из N состояний и последовательности длиной T.

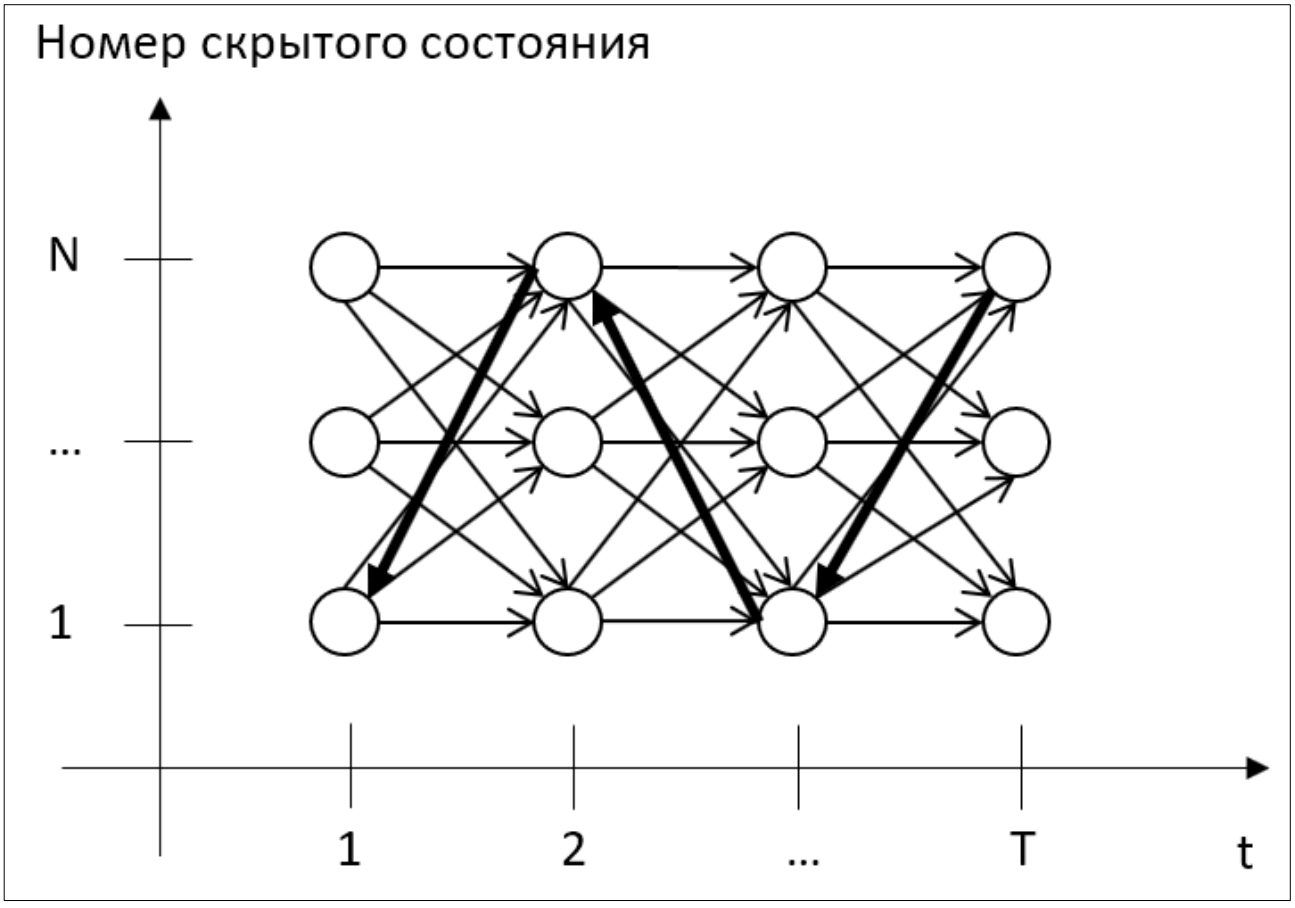

Рисунок 1 – Иллюстрация работы алгоритма Витерби

По оси абсцисс отложены моменты времени, а по оси ординат – номера скрытых состояний, в которых может быть процесс, описываемый моделью. Тонкими стрелками показано направление вычислений, благодаря которым алгоритм Витерби исследует все возможные последовательности скрытых состояний во время прямого прохода, а полужирными стрелками показан наиболее вероятный путь, который выбирает алгоритм на этапе рекурсивного определения наиболее вероятной последовательности скрытых состояний.

1.4 Распознавание последовательностей, описываемых скрытыми марковскими моделями

Пусть определено несколько классов, соответствующих некоторым различным случайным процессам с номерами $\overline{1,D}$, которые описываются соответствующими СММ $\lambda_1, ..., \lambda_D$, а также имеется последовательность многомерных наблюдений $O = \{\boldsymbol{o}_1, ..., \boldsymbol{o}_T\}$. Для классификации последовательности, т.е. определения того, каким именно процессом, описываемым соответствующей СММ, она была порождена, как правило, применяют критерий максимума функции правдоподобия (МФП). В этом случае последовательность O относят к тому классу r^*, для которого значение функции правдоподобия является максимальным: $r^* = \arg\max_{r \in 1,...D} \left(p\left(O \mid \lambda_r\right)\right)$.

Для расчёта значения функции правдоподобия того, что последовательность O была сгенерирована процессом, описываемым СММ λ, т. е. $p\left(O \mid \lambda\right) = \sum_{q_1, q_2, ..., q_T} p\left(\{\boldsymbol{o}_1, ..., \boldsymbol{o}_T\}, \{q_1, q_2, ..., q_T\} \mid \lambda\right)$ обычно применяют алгоритм forward-backward (прямой-обратный) [24, 63]. Для вычисления самого значения функции правдоподобия $L = p\left(O \mid \lambda\right)$ необходима лишь прямая часть forward-backward алгоритма, однако для полноты далее приводится и обратная часть алгоритма, так как она пригодится в дальнейшем для описания алгоритма обучения.

Первая часть forward-backward алгоритма производит вычисление прямых вероятностей $\alpha_t(i) = p(\{\boldsymbol{o}_1, \boldsymbol{o}_2, ..., \boldsymbol{o}_t\}, q_t = s_i \mid \lambda)$, $t = \overline{1,T}$, $i = \overline{1,N}$, т. е. вероятностей того, что последовательность многомерных наблюдений $\{\boldsymbol{o}_1, \boldsymbol{o}_2, ..., \boldsymbol{o}_t\}$ была порождена процессом, описываемым моделью λ, и что данный процесс находился в скрытом состоянии s_i в момент времени t. Алгоритм расчёта прямых вероятностей и значения функции правдоподобия:

1) инициализация:

$$\alpha_1(i) = \pi_i b_i(\boldsymbol{o}_1), \; i = \overline{1,N}; \tag{7}$$

2) индукция:

$$\alpha_{t+1}(j) = \left[\sum_{i=1}^{N} \alpha_t(i) a_{ij} \right] b_j(\boldsymbol{o}_{t+1}), \; i = \overline{1,N}, \; t = \overline{1,T-1}; \tag{8}$$

3) завершение:

$$p(O \mid \lambda) = \sum_{i=1}^{N} \alpha_T(i). \tag{9}$$

Рисунок 2 содержит иллюстрацию работы алгоритма вычисления прямых вероятностей для СММ, состоящей из N состояний и последовательности длиной T.

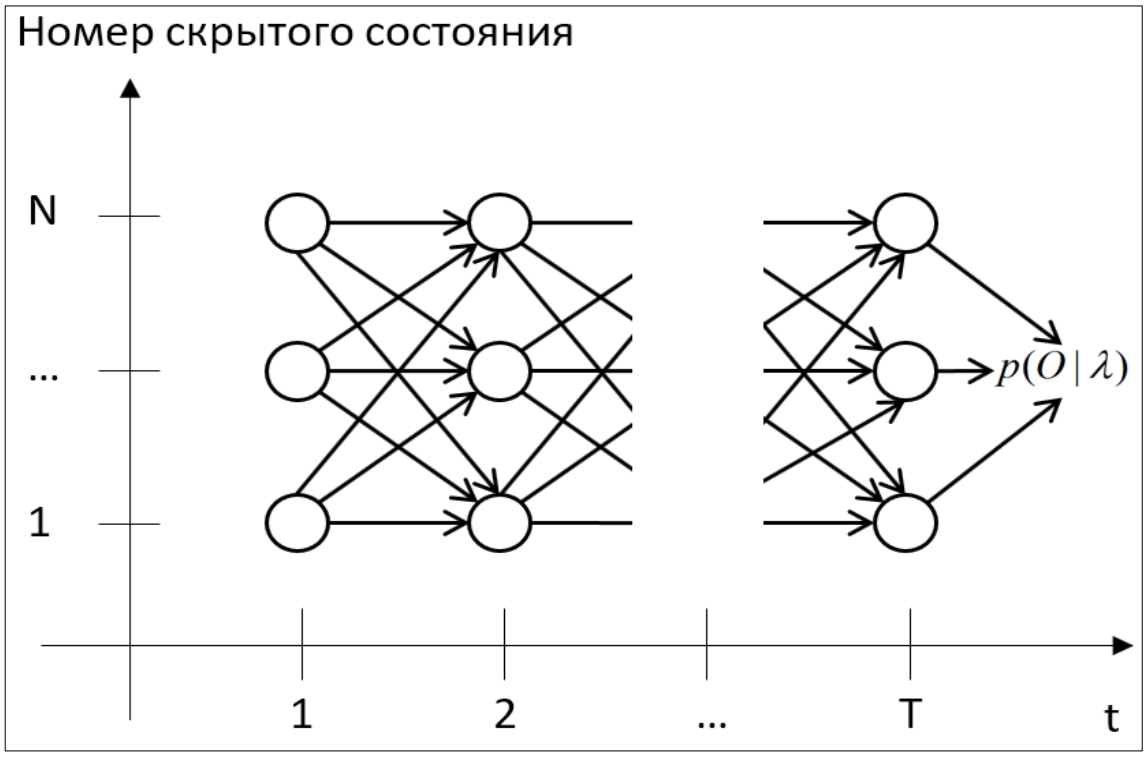

Рисунок 2 – Иллюстрация работы алгоритма вычисления прямых вероятностей

По оси абсцисс отложены моменты времени, а по оси ординат – номера скрытых состояний, в которых может быть процесс, описываемый моделью. Стрелки показывают направление вычислений, а круги обозначают прямые вероятности.

Вторая часть forward-backward алгоритма позволяет рассчитать обратные вероятности (backward variables) $\beta_t(i) = p(\{\boldsymbol{o}_{t+1}, \boldsymbol{o}_{t+2}, ..., \boldsymbol{o}_T\} \mid q_t = s_i, \lambda), t = \overline{1,T},\ i = \overline{1,N}$, т. е. вероятности того, что модель λ в момент времени t находилась в состоянии s_i, а затем описываемым ей процессом была порождена последовательность наблюдений $\{\boldsymbol{o}_{t+1}, \boldsymbol{o}_{t+1}, \ldots, \boldsymbol{o}_T\}$. Алгоритм вычисления обратных вероятностей:

1) инициализация:

$$\beta_T(i) = 1, \quad i = \overline{1,N}; \tag{10}$$

2) индукция:

$$\beta_t(i) = \sum_{j=1}^{N} a_{ij} b_j(\boldsymbol{o}_{t+1}) \beta_{t+1}(j),\ i = \overline{1,N},\ t = \overline{1,T-1}; \tag{11}$$

3) завершение (приведено для общности):

$$p(O \mid \lambda) = \sum_{i=1}^{N} \beta_1(i)\, \pi_i\, b_i(\boldsymbol{o}_1). \tag{12}$$

Рисунок 3 содержит иллюстрацию работы алгоритма вычисления обратных вероятностей для СММ, состоящей из N состояний и последовательности длиной T. По оси абсцисс отложены моменты времени, а по оси ординат – номера скрытых состояний, в которых может находиться процесс, описываемый моделью. Стрелки показывают направление вычислений, а круги обозначают обратные вероятности.

Таким образом, после рекурсивного вычисления прямых вероятностей по формулам (7)-(8), формула (9) позволяет вычислить искомое значение функции правдоподобия того, что последовательность O была порождена процессом, описываемым СММ λ. Также для общности показано, что значение этой функции также можно вычислить и с помощью обратных вероятностей по формуле (12).

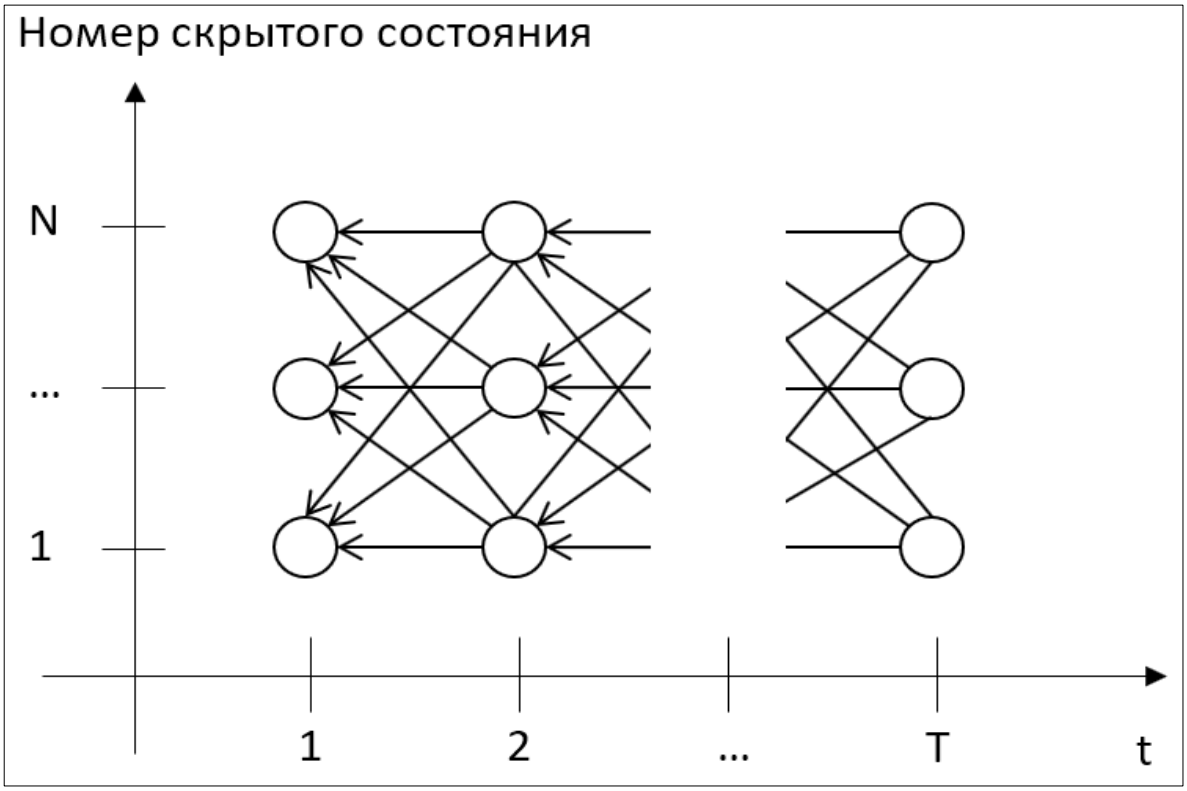

Рисунок 3 – Иллюстрация работы алгоритма вычисления обратных вероятностей

Вообще говоря, точность распознавания последовательностей будет зависеть от многих параметров. Чем больше конкурирующих классов случайных процессов, чем более они схожи по параметрам, чем короче обрабатываемые последовательности, тем больше ошибок будет допускаться при классификации.

1.5 Обучение скрытой марковской модели

1.5.1 Алгоритм Баума-Велша

Для представления по имеющимся последовательностям изучаемого случайного процесса в виде СММ необходимо найти оценку параметров этой модели. Для этого необходимо решить задачу обучения, которая состоит в настройке параметров модели $\lambda = \{\Pi, A, B\}$ (т.е. распределения начального скрытого состояния, матрицы переходов и распределения наблюдений в скрытых состояниях) по последовательности наблюдений $O^* = \{O^1, O^2, ..., O^K\}$, где K – это число наблюдаемых

последовательностей. Для решения этой задачи, как правило, применяется метод обучения, использующий процедуру максимизации функции правдоподобия:

$$L(O^* \mid \lambda) = \prod_{k=1}^{K} p(O^k \mid \lambda) \cdot \tag{13}$$

Для этой процедуры известен эффективный алгоритм Баума-Велша [24, 60], который является частным случаем алгоритма EM (EM – expectation-maximization; ожидание-максимизация) [64-65]. Так как алгоритм является итеративным, то перед началом его работы необходимо задать некоторое начальное приближение λ параметров СММ.

Опишем далее вычисление вспомогательных величин для одной из обучающих последовательностей O. Для более краткого описания алгоритма Баума-Велша введем вероятности γ, ξ:

$$\gamma_t(i) = p(q_t = s_i \mid O, \lambda) = \frac{\alpha_t(i)\beta_t(i)}{p(O \mid \lambda)}, \quad i = \overline{1,N}, \; t = \overline{1,T}, \tag{14}$$

$$\begin{gathered} \xi_t(i,j) = p(q_t = s_i, q_{t+1} = s_j \mid O, \lambda) = \\ = \frac{\alpha_t(i) a_{ij} b_j(\boldsymbol{o}_{t+1}) \beta_{t+1}(j)}{p(O \mid \lambda)}, \quad i, j = \overline{1,N}, \quad t = \overline{1,T-1} \end{gathered}, \tag{15}$$

для СММ с непрерывной плотностью распределения наблюдений также введём вероятность:

$$\gamma_t(i,m) = p(q_t = i, \omega_{it} = m \mid O, \lambda) = \gamma_t(i) \left[\frac{\tau_{im} g(\boldsymbol{o}_t, \mu_{im}, \Sigma_{im})}{b_i(\boldsymbol{o}_t)} \right], \tag{16}$$

где $\hat{\lambda}$ – текущая оценка параметров модели, а ω_{it} – компонента смеси нормальных распределений в момент времени t для состояния i. Отметим, что в формулах (14)-(15) задействуются прямые и обратные вероятности, которые вычисляются с использованием алгоритма forward-backward по формулам (7)-(11). Кроме того, следует заметить, что для каждой обучающей последовательности под индексом $k = \overline{1,K}$ вычисляется свой набор значений прямых и обратных вероятностей, а

также вероятностей γ, ξ. Для их обозначения используется соответствующий индекс: $\alpha^{(k)}, \beta^{(k)}, \gamma^{(k)}, \xi^{(k)}$.

С учетом введенных обозначений для СММ с непрерывным распределением наблюдений новое приближение оценок параметров будет находиться в точке $\hat{\lambda}'$ с координатами [66]:

$$\hat{\pi}'_i = \frac{1}{K}\sum_{k=1}^{K}\gamma_1^{(k)}(i), \tag{17}$$

$$\hat{a}'_{ij} = \frac{\sum_{k=1}^{K}\sum_{t=1}^{T^k-1}\xi_t^{(k)}(i,j)}{\sum_{k=1}^{K}\sum_{t=1}^{T^k-1}\gamma_t^{(k)}(i)}, \tag{18}$$

$$i, j = \overline{1, N}$$

для СММ с дискретным распределением наблюдений:

$$\hat{b}'_i(v_m) = \frac{\sum_{k=1}^{K}\sum_{t=1}^{T^k}\gamma_t^{(k)}(i)}{\sum_{k=1}^{K}\sum_{t=1}^{T^k}\gamma_t^{(k)}(i)}, \quad npu\ O_t = v_m \tag{19}$$

$$i = \overline{1, N}, \quad m = \overline{1, M}$$

для СММ с непрерывной плотностью распределения наблюдений:

$$\hat{\tau}'_{im} = \frac{\sum_{k=1}^{K}\sum_{t=1}^{T^k-1}\gamma_t^{(k)}(i,m)}{\sum_{k=1}^{K}\sum_{t=1}^{T^k-1}\gamma_t^{(k)}(i)}, \tag{20}$$

$$\hat{\mu}'_{im} = \frac{\sum_{k=1}^{K}\sum_{t=1}^{T^k-1}\gamma_t^{(k)}(i,m)\,\boldsymbol{o}_t^k}{\sum_{k=1}^{K}\sum_{t=1}^{T^k-1}\gamma_t^{(k)}(i,m)}, \tag{21}$$

$$\hat{\Sigma}'_{im} = \frac{\sum_{k=1}^{K}\sum_{t=1}^{T^k-1}\gamma_t^{(k)}(i,m)(\boldsymbol{o}_t^{(k)} - \hat{\mu}'_{im})(\boldsymbol{o}_t^{(k)} - \hat{\mu}'_{im})^T}{\sum_{k=1}^{K}\sum_{t=1}^{T^k-1}\gamma_t^{(k)}(i,m)}, \qquad (22)$$

$$i = \overline{1,N}, \quad m = \overline{1,M}.$$

С помощью выражений (17)-(22) выполняется итерационное улучшение оценок параметров СММ. При этом на каждой новой итерации производится перерасчёт переменных γ, ξ по формулам (14)-(16) с параметрами $\hat{\lambda} = \hat{\lambda}'$. Леонард Баум и его коллеги доказали, что получаемая оценка модели $\hat{\lambda}'$ будет более правдоподобной, т.е., что $L\left(O^*\middle|\hat{\lambda}'\right) \geq L\left(O^*\middle|\hat{\lambda}\right)$ [1]. Алгоритм Баума-Велша в общем случае не обязательно сходится к глобальному максимуму, поэтому рекомендуется запускать его поочередно на нескольких различных начальных приближениях параметров, выбирая в итоге наилучшую оценку. Данный процесс итеративного улучшения оценок следует продолжать до тех пор, пока не выполнится некоторое условие останова. К примеру, можно следить за абсолютной разностью между логарифмами функции правдоподобия, вычисляемыми по формуле (13), вычисленными на текущем и предыдущих шагах, и, если она меньше заранее заданного малого числа ε, производить выход из итеративного цикла.

Выбор же таких характеристик модели как количество скрытых состояний, количество символов в алфавите или количество компонент смесей проводится исходя из специфики решаемой задачи. Выбор количества символов в алфавите проводится исходя количества уникальных дискретных наблюдений, порождаемых описываемым процессом. Остальные характеристики выбираются в зависимости от требуемой точности решения задачи. Тем не менее, если у описываемого процесса удаётся выделить состояния, интерпретируемые в предметной области, то количество скрытых состояний модели разумно принять равным количеству данных состояний.

1.5.2 Генерация начальных приближений параметров скрытой марковской модели

В случае СММ с дискретным распределением наблюдений при построении начальных приближений параметров модели значения $\{\Pi, A, B\}$ можно задавать случайным образом, т.е. генерировать с помощью равномерного распределения компоненты вектора Π, компоненты матрицы A и вероятности условного дискретного распределения B, так, чтобы они удовлетворяли вероятностным ограничениям на параметры модели [24].

В случае же СММ с непрерывной плотностью распределения наблюдений, параметры Π и A модели можно генерировать аналогично, однако генерация параметров условных плотностей распределения наблюдений B, описываемых смесями нормальных распределений, представляет значительно большую сложность. Как показывает практика, для успешного обучения СММ с непрерывным распределением критически важно подобрать правдоподобные значения весов смесей τ_{im}, $i=\overline{1,N}, m=\overline{1,M}$, векторов математических ожиданий μ_{im}, $i=\overline{1,N}, m=\overline{1,M}$ и ковариационных матриц Σ_{im}, $i=\overline{1,N}, m=\overline{1,M}$. Автор данной диссертационной работы считает, что в данной ситуации наиболее рациональным (при отсутствии априорных знаний о распределении наблюдений) использовать модель смесей нормальных распределений (Gaussian mixture model – GMM) [67] для оценивания параметров распределения, так как такие же смеси используются и самой СММ.

Модель GMM описывает исходную выборку X исходя из предположения, что она была сгенерирована смесью нормальных распределений: $p(\boldsymbol{x}\,|\,\theta)=\sum_{m=1}^{\tilde{M}}\tilde{\tau}_m\, g(\boldsymbol{x};\tilde{\mu}_m,\tilde{\Sigma}_m)$, $m=\overline{1,\tilde{M}}$, $\boldsymbol{x}\in R^Z$, где $\boldsymbol{x}$ - Z-мерное наблюдение из выборки X, $\tilde{\tau}_m$ – вес m-й компоненты смеси, $\tilde{\mu}_m$ – вектор математического ожидания нормального распределения, соответствующего m-й компоненте смеси, $\tilde{\Sigma}_m$ – ковариационная матрица нормального распределения, соответствующая m-й

компоненте смеси, $g(\boldsymbol{x};\mu,\Sigma)$ – функция плотности многомерного нормального распределения, а $\theta=\left\{\theta_m=\left(\tilde{\tau}_m,\tilde{\mu}_m,\tilde{\Sigma}_m\right),m=\overline{1,\tilde{M}}\right\}$ - набор параметров GMM. Оценка параметров GMM находится путём максимизации функции правдоподобия вида $L(\theta;X)=\prod_{x\in X}p(x\mid\theta)$ при помощи EM-алгоритма.

Для генерации начального приближения CMM предлагается следующий алгоритм:

1) найти оценку параметров GMM, имеющей N (количество скрытых состояний CMM) компонент смесей, по выборке наблюдений, принадлежащих обучающим последовательностям из множества 0^*, получив в результате N оценок компонент смесей GMM: $\left\{\tilde{\tau}_i,\tilde{\mu}_i,\tilde{\Sigma}_i\mid i=\overline{1,N}\right\}$ и разделив всё множество наблюдений на N групп, отнеся каждое наблюдение $\boldsymbol{o}$ к группе $\tilde{O}_i, i=\overline{1,N}$, $i=\arg\max_{i=\overline{1,N}}\tilde{\tau}_i g(\boldsymbol{o};\tilde{\mu}_i,\tilde{\Sigma}_i)$;

2) для каждой группы с номером $i=\overline{1,N}$ найти оценку параметров GMM, имеющей M (количество компонент смесей CMM) компонент смесей, по выборке наблюдений, принадлежащих $\tilde{O}_i$, получив в результате $N\times M$ оценок компонент смесей GMM: $\left\{\tilde{\tau}_{im},\tilde{\mu}_{im},\tilde{\Sigma}_{im}\mid i=\overline{1,N},m=\overline{1,M}\right\}$;

3) В качестве параметра B начального приближения CMM использовать найденные оценки GMM: $B^0=\left\{\tilde{\tau}_{im},\tilde{\mu}_{im},\tilde{\Sigma}_{im}\mid i=\overline{1,N},m=\overline{1,M}\right\}$.

Также можно добавить небольшой нормально распределенный шум к параметрам $\mu_{im},\Sigma_{im},i=\overline{1,N},m=\overline{1,M}$, чтобы эти компоненты начальных приближений различались.

1.6 Масштабирование вероятностей в формулах анализа последовательностей, описываемых скрытыми марковскими моделями

Ранее во всех формулах вероятности умножались друг на друга, то есть числа меньше единицы, имеющие, как правило, значения обратные количеству скрытых состояний, умножаются в количестве, пропорциональном длине последовательности. Для длинных последовательностей (длиной более 100) данные произведения достаточно быстро становятся меньше минимальных аппаратно реализуемых чисел современных машин. Для исправления данной проблемы необходимо либо использовать длинную арифметику, что значительно замедлит вычисления, либо масштабировать все промежуточные произведения, чтобы они не стремились к нулю.

1.6.1 Масштабирование вычислений в алгоритме Витерби

Идея масштабированного алгоритма Витерби заключается в использовании логарифмов при вычислении вероятностей. Приведём шаги данной версии алгоритма далее [24].

Масштабированный алгоритм Витерби:

1) инициализация:

$$\tilde{\delta}_1(i) = \ln \pi_i + \ln b_i(\boldsymbol{o}_1), \quad i = \overline{1, N}, \tag{23}$$

$$\tilde{\psi}_1(i) = 0, \quad i = \overline{1, N}; \tag{24}$$

2) индукция:

$$\tilde{\delta}_t(j) = \max_{1 \le i \le N} \left[\tilde{\delta}_{t-1}(i) + \ln a_{ij} \right] + \ln b_j(\boldsymbol{o}_t), \quad j = \overline{1, N}, \quad t = \overline{2, T} \tag{25}$$

$$\tilde{\psi}_t(j) = \arg\max_{1 \le i \le N} \left[\tilde{\delta}_{t-1}(i) + \ln a_{ij} \right], \quad j = \overline{1, N}, \ t = \overline{2, T} \tag{26}$$

3) завершение:

$$\hat{q}_T = \arg\max_{1 \le i \le N} \left[\tilde{\delta}_T(i) \right], \tag{27}$$

4) рекурсивное определение наиболее вероятной последовательности скрытых состояний:

$$\hat{q}_t = \tilde{\psi}_{t+1}(\hat{q}_t), \quad t = \overline{T-1, \ 1}. \tag{28}$$

После завершения алгоритма получим сформированную наиболее вероятную последовательность скрытых состояний: $\hat{Q} = \{\hat{q}_1, \cdots, \hat{q}_T\}$.

1.6.2 Масштабирование вычислений в алгоритме forward-backward

Далее приведён один из наиболее эффективных методов масштабирования прямых и обратных вероятностей, который практически не замедляет обучение [24, 68].

Введём параметр масштаба для прямых вероятностей $C_t = \dfrac{1}{\sum\limits_{i=1}^{N} \alpha_t(i)}$, $t = \overline{1,T}$, а также обозначим отмасштабированные прямые вероятности как $\hat{\alpha}_t(i) = C_t \alpha_t(i) = \dfrac{\alpha_t(i)}{\sum\limits_{i=1}^{N} \alpha_t(i)}$, $i = \overline{1,N}, t = \overline{1,T}$. Для вычисления отмасштабированных прямых вероятностей используем следующие выражения:

1) инициализация:

$$\overline{\alpha}_1(i) = \alpha_1(i), \qquad i = \overline{1,N} \tag{29}$$

$$c_1 = \frac{1}{\sum\limits_{i=1}^{N} \alpha_1(i)}, \tag{30}$$

$$\hat{\alpha}_1(i) = C_1 \alpha_1(i) = \frac{\alpha_1(i)}{\sum\limits_{i=1}^{N} \alpha_1(i)} = c_1 \overline{\alpha}_1(i), \qquad i = \overline{1,N}; \tag{31}$$

2) индукция:

$$\overline{\alpha}_{t+1}(j) = C_t \alpha_{t+1}(j) = \left[\sum_{i=1}^{N} C_t \alpha_t(i) a_{ij}\right] b_j(\boldsymbol{o}_{t+1}) = \left[\sum_{i=1}^{N} \hat{\alpha}_t(i) a_{ij}\right] b_j(\boldsymbol{o}_{t+1}), \quad \begin{matrix} j = \overline{1,N} \\ t = \overline{1,T-1} \end{matrix}, \tag{32}$$

$$c_{t+1} = \frac{1}{C_t \sum\limits_{i=1}^{N} \alpha_{t+1}(i)} = \frac{1}{\sum\limits_{i=1}^{N} \overline{\alpha}_{t+1}(i)}, \quad t = \overline{1,T-1}, \tag{33}$$

$$\hat{\alpha}_{t+1}(i) = C_t \alpha_t(i) = \frac{\alpha_{t+1}(i)}{\sum_{i=1}^{N} \alpha_{t+1}(i)} = \frac{1}{C_t \sum_{i=1}^{N} \alpha_{t+1}(i)} C_t \alpha_{t+1}(i) = c_{t+1} \overline{\alpha}_{t+1}(i), \quad \begin{matrix} j = \overline{1,N} \\ t = \overline{1,T-1} \end{matrix} . \quad (34)$$

Заметим, что $C_t = \frac{1}{c_{t+1}} \frac{1}{\sum_{i=1}^{N} \alpha_{t+1}(i)} = \frac{C_{t+1}}{c_{t+1}}$ и, следовательно, $C_t = C_{t-1} c_t = \prod_{\tau=1}^{t} c_\tau$.

Введём параметр масштаба для обратных вероятностей $D_t = \prod_{\tau=t}^{T} c_\tau, \ t = \overline{1,T}$, а также обозначим отмасштабированные обратные вероятности как $\hat{\beta}_t(i) = D_t \beta_t(i), \ \ i = \overline{1,N}, t = \overline{1,T}$. Для вычисления отмасштабированных обратных вероятностей используем следующие выражения:

1) инициализация:

$$\hat{\beta}_t(i) = D_t \beta_T(i) = c_T \beta_T(i), \qquad i = \overline{1,N}, \ t = \overline{1,T}; \quad (35)$$

2) индукция:

$$\overline{\beta}_t(i) = D_{t+1} \beta_t(i) = \sum_{j=1}^{N} a_{ij} b_j(\boldsymbol{o}_{t+1}) D_{t+1} \beta_{t+1}(j) = \sum_{j=1}^{N} a_{ij} b_j(\boldsymbol{o}_{t+1}) \hat{\beta}_{t+1}(j), \quad \begin{matrix} j = \overline{1,N} \\ t = \overline{1,T-1} \end{matrix} , \quad (36)$$

$$\hat{\beta}_t(i) = D_t \beta_t(i) = c_t D_{t+1} \beta_t(i) = c_t \overline{\beta}_t(i), \qquad i = \overline{1,N}, \ t = \overline{1,T} . \quad (37)$$

Отметим, что:

$$P(O \mid \lambda) = \sum_{i=1}^{N} \alpha_T(i) = \frac{1}{C_T},$$

а также, что:

$$C_t D_{t+1} = \prod_{\tau=1}^{t} c_\tau \prod_{\tau=t+1}^{T} c_\tau = \prod_{\tau=1}^{T} c_\tau = C_T, \quad t = \overline{1,T-1}.$$

Так как значение функции правдоподобия (13) в случае длинных последовательностей также будет слишком мало для машинного представления, рекомендуется вычислять вместо неё её логарифм:

$$\ln L(O^* \mid \lambda) = \ln \prod_{k=1}^{K} p(O^k \mid \lambda) = \ln \prod_{k=1}^{K} \frac{1}{C^T} = \ln \prod_{k=1}^{K} \left(\prod_{t=1}^{T} c_t \right)^{-1} = -\sum_{k=1}^{K} \sum_{t=1}^{T} \ln c_t^{(k)} . \quad (38)$$

1.6.3 Масштабирование вычислений в алгоритме Баума-Велша

Далее покажем, как отмасштабированные прямые и обратные вероятности используются для масштабирования вычисления в алгоритме Баума-Велша [24, 68]. Подставим $\frac{\hat{\alpha}_t(i)}{C_t}$ вместо $\alpha_t(i)$ и $\frac{\hat{\beta}_t(i)}{D_t}$ вместо $\beta_t(i)$, $i=\overline{1,N}, t=\overline{1,T}$ в формулах (14) и (15) :

$$\gamma_t(i)=\frac{\alpha_t(i)\beta_t(j)}{p(O\mid\lambda)}=\frac{\frac{\hat{\alpha}_t(i)}{C_t}\frac{\hat{\beta}_t(j)}{D_t}}{\frac{1}{C_T}}=\frac{\hat{\alpha}_t(i)\hat{\beta}_t(j)\frac{1}{c_tC_T}}{\frac{1}{C_T}}=\frac{\hat{\alpha}_t(i)\hat{\beta}_t(j)}{c_t},\quad \begin{matrix} i=\overline{1,N} \\ t=\overline{1,T} \end{matrix}, \qquad (39)$$

$$\xi_t(i,j)=\frac{\alpha_t(i)a_{ij}b_j(o_{t+1})\beta_{t+1}(j)}{P(O\mid\lambda)}=\frac{\frac{\hat{\alpha}_t(i)}{C_t}a_{ij}b_j(o_{t+1})\frac{\hat{\beta}_{t+1}(j)}{D_{t+1}}}{\frac{1}{C_T}}=$$

$$=\frac{\hat{\alpha}_t(i)a_{ij}b_j(o_{t+1})\hat{\beta}_{t+1}(j)\frac{1}{C_T}}{\frac{1}{C_T}}=\hat{\alpha}_t(i)a_{ij}b_j(o_{t+1})\hat{\beta}_{t+1}(j), \qquad (40)$$

$$t=\overline{1,T-1},\ i,j=\overline{1,N}.$$

Далее выражения (39) и (40), использующие отмасштабированные прямые и обратные вероятности, следует использовать вместо выражений (14) и (15) соответственно при вычислении формул (16)-(22), использующихся в алгоритме Баума-Велша.

1.7 Вычисление первых производных от логарифма функции правдоподобия того, что описываемый скрытой марковской моделью процесс сгенерировал последовательность

Приведем далее способ вычисления производных от логарифма функции правдоподобия по параметрам СММ $\frac{\partial \ln L(O^* \mid \lambda)}{\partial \eta}$, где η - некоторый параметр СММ, для СММ с непрерывной плотностью распределения наблюдений [69-70].

Принимая во внимание выражение (38), проведём следующее преобразование:

$$\frac{\partial P\left(O|\lambda\right)}{\partial \eta} = -\left(\prod_{t=1}^{T} c_t\right)^{-2} \frac{\partial\left(\prod_{t=1}^{T} c_t\right)}{\partial \eta} = -\left(\prod_{t=1}^{T} c_t\right)^{-2} \prod_{t=1}^{T} c_t \left(\sum_{t=1}^{T} \frac{1}{c_t} \frac{\partial c_t}{\partial \eta}\right) = \tag{41}$$

$$= -\left(\prod_{t=1}^{T} c_t\right)^{-1} \left(\sum_{t=1}^{T} \frac{1}{c_t} \frac{\partial c_t}{\partial \eta}\right) = -P(O \mid \lambda) \left(\sum_{t=1}^{T} \frac{1}{c_t} \frac{\partial c_t}{\partial \eta}\right).$$

Следовательно:

$$\frac{\partial \ln L(O^* \mid \lambda)}{\partial \eta} = \sum_{k=1}^{K} P\left(O^k \mid \lambda\right)^{-1} \frac{\partial P\left(O^k \mid \lambda\right)}{\partial \eta} = -\sum_{k=1}^{K} \left(\sum_{t=1}^{T} \frac{1}{c_t^k} \frac{\partial c_t^k}{\partial \eta} \right) . \tag{42}$$

Вычисление производной от параметра масштаба по некоторому параметру модели η производится следующим образом:

$$\frac{\partial c_t}{\partial \eta} = \frac{\partial}{\partial \eta}\left(\left(\sum_{i=1}^{N} \overline{\alpha}_t(i)\right)^{-1}\right) = -\left(\sum_{i=1}^{N} \overline{\alpha}_t(i)\right)^{-2} \sum_{i=1}^{N} \frac{\partial \overline{\alpha}_t(i)}{\partial \eta} = -c_t^2 \sum_{i=1}^{N} \frac{\partial \overline{\alpha}_t(i)}{\partial \eta}, t = \overline{1,T} . \tag{43}$$

Для вычисления $\frac{\partial \overline{\alpha}_t(i)}{\partial \eta}, i = \overline{1,N}$ продифференцируем по шагам алгоритм вычисления промежуточных прямых переменных с масштабом:

1 шаг.

$$\frac{\partial \overline{\alpha}(i)}{\partial \eta} = \frac{\partial \alpha_1(i)}{\partial \eta}, \quad i = \overline{1,N} . \tag{44}$$

2 шаг.

$$\frac{\partial\overline{\alpha}_t(i)}{\partial\eta}=\frac{\partial}{\partial\eta}\left(\left[\sum_{j=1}^{N}\hat{\alpha}_{t-1}(j)a_{ji}\right]b_i(\boldsymbol{o}_t)\right)=$$
$$=\left[\sum_{j=1}^{N}\left(\frac{\partial\hat{\alpha}_{t-1}(j)}{\partial\eta}a_{ji}+\hat{\alpha}_{t-1}(j)\frac{\partial a_{ji}}{\partial\eta}\right)\right]b_i(\boldsymbol{o}_t)+\sum_{j=1}^{N}\left(\hat{\alpha}_{t-1}(j)a_{ji}\right)\frac{\partial b_i(\boldsymbol{o}_t)}{\partial\eta}, \quad (45)$$

где $\frac{\partial\hat{\alpha}_{t-1}(j)}{\partial\eta}=\frac{\partial c_{t-1}}{\partial\eta}\overline{\alpha}_{t-1}(j)+\frac{\partial\overline{\alpha}_{t-1}(j)}{\partial\eta}c_{t-1},\ i=\overline{1,N},\ t=\overline{2,T}$.

Таким образом, для вычисления значений $\frac{\partial\overline{\alpha}_t(i)}{\partial\eta}, i=\overline{1,N}$ нам потребуется вычислить производные $\frac{\partial\overline{\alpha}_1(i)}{\partial\eta},\frac{\partial b_i(t)}{\partial\eta},\frac{\partial a_{ij}}{\partial\eta},\quad i,j=\overline{1,N}, t=\overline{1,T}$, где

$$\overline{\alpha}_1(i)=\alpha_1(i)=\pi_i b_i(\boldsymbol{o}_1),\quad i=\overline{1,N},$$

$$b_i(t)=\sum_{m=1}^{M}\tau_{im}g(\boldsymbol{o}_t;\mu_{im},\Sigma_{im}),\ i=\overline{1,N},\ \ t=\overline{1,T}, \text{ где} \quad (46)$$

$$g(\boldsymbol{o}_t;\mu_{im},\Sigma_{im})=\frac{1}{\sqrt{(2\pi)^Z\left|\Sigma_{im}\right|}}e^{-0.5*(\boldsymbol{o}_t-\mu_{im})^T\Sigma_{im}^{-1}(\boldsymbol{o}_t-\mu_{im})}, \quad (47)$$
$$t=\overline{1,T},\ i=\overline{1,N},\ m=\overline{1,M},$$

а число Z — размерность наблюдений.

В случае недиагональной матрицы при вычислении производной по элементу ковариационной матрицы придется дифференцировать элементы обратной матрицы, поэтому будем рассматривать случай, когда матрицы $\Sigma_{im},\ i=\overline{1,N},\ \ m=\overline{1,M}$ являются диагональными.

Выпишем способ вычисления производных $\frac{\partial\overline{\alpha}_1(i)}{\partial\eta},\frac{\partial b_i(t)}{\partial\eta},\frac{\partial a_{ij}}{\partial\eta}$, $i,j=\overline{1,N}, t=\overline{1,T}$ для указанных значений параметра η (параметр η может принимать значения $\pi_i, a_{ij}, \tau_{im}, \mu_{im}^z, \Sigma_{im}^{zz},\ \ i,j=\overline{1,N}, m=\overline{1,M}, z=\overline{1,Z}$):

$$\frac{\partial\overline{\alpha}_1(i)}{\partial\pi_j}=\begin{cases}b_i(\boldsymbol{o}_1),\ i=j\\0,\quad i\neq j\end{cases},\quad i,j=\overline{1,N}, \quad (48)$$

$$\frac{\partial b_i(\boldsymbol{o}_t)}{\partial \pi_j} = 0, \quad i, j = \overline{1,N}, \quad t = \overline{1,T}, \tag{49}$$

$$\frac{\partial \overline{\alpha}_1(i)}{\partial a_{i_1 j_1}} = 0, \ i, i_1, j_1 = \overline{1,N}, \tag{50}$$

$$\frac{\partial b_i(\boldsymbol{o}_t)}{\partial a_{i_1 j_1}} = 0, \quad i, i_1, j_1 = \overline{1,N}, \ t = \overline{1,T}, \tag{51}$$

$$\frac{\partial a_{ij}}{\partial x} = \begin{cases} 1, & x = a_{ij} \\ 0, & x \neq a_{ij} \end{cases}, \quad i, j = \overline{1,N}, \tag{52}$$

$$\frac{\partial b_i(\boldsymbol{o}_t)}{\partial \tau_{i_1 m}} = \begin{cases} g(\boldsymbol{o}_t; \mu_{im}, \Sigma_{im}), & i = i_1 \\ 0, & i \neq i_1 \end{cases}, \quad i, i_1 = \overline{1,N}, \ t = \overline{1,T}, \ m = \overline{1,M}, \tag{53}$$

$$\frac{\partial \overline{\alpha}_1(i)}{\partial \tau_{i_1 m}} = \begin{cases} \pi_i \dfrac{\partial b_i(\boldsymbol{o}_1)}{\partial \tau_{i_1 m}}, & i = i_1 \\ 0, & i \neq i_1 \end{cases}, \quad i, i_1 = \overline{1,N}, \ m = \overline{1,M}, \tag{54}$$

$$\frac{\partial b_i(\boldsymbol{o}_t)}{\partial \mu_{i_1 m}^z} = \begin{cases} 0.5\tau_{im} g(\boldsymbol{o}_t; \mu_{im}, \Sigma_{im}) \dfrac{\boldsymbol{o}_t^z - \mu_{im}^z}{\Sigma_{im}^{zz}}, & i = i_1 \\ 0, & i \neq i_1 \end{cases}, \tag{55}$$

$$i, i_1 = \overline{1,N}, \ t = \overline{1,T}, \ m = \overline{1,M}, \quad z = \overline{1,Z}$$

$$\frac{\partial \overline{\alpha}_1(i)}{\partial \mu_{i_1 m}^z} = \begin{cases} \pi_i \dfrac{\partial b_i(\boldsymbol{o}_1)}{\partial \mu_{i_1 m}^z}, & i = i_1 \\ 0, & i \neq i_1 \end{cases}, \quad i, i_1 = \overline{1,N}, \ m = \overline{1,M}, \quad z = \overline{1,Z}, \tag{56}$$

$$\frac{\partial b_i(t)}{\partial \Sigma_{i_1 m}^{zz}} = \begin{cases} 0.5\tau_{im} g(\boldsymbol{o}_t; \mu_{im}, \Sigma_{im}) \left(\left(\dfrac{\boldsymbol{o}_t^z - \mu_{im}^z}{\Sigma_{im}^{zz}} \right)^2 - \dfrac{1}{\left| \Sigma_{im} \right|} \right), & i = i_1 \\ 0, & i \neq i_1 \end{cases}, \tag{57}$$

$$i, i_1 = \overline{1,N}, \ t = \overline{1,T}, \ m = \overline{1,M}, \quad z = \overline{1,Z}$$

$$\frac{\partial \overline{\alpha}_1(i)}{\partial \Sigma_{i_1 m}^{zz}} = \begin{cases} \pi_i \dfrac{\partial b_i(\boldsymbol{o}_1)}{\partial \Sigma_{im}^{zz}}, & i = i_1 \\ 0, & i \neq i_1 \end{cases}, \ i, i_1 = \overline{1,N}, \ m = \overline{1,M}, \quad z = \overline{1,Z}. \tag{58}$$

1.8 Моделирование последовательностей, описываемых скрытыми марковскими моделями

1.8.1 Моделирование целых последовательностей

Моделирование целых последовательностей, описываемых СММ с параметрами $\lambda = (\Pi, A, B)$, производится по следующему алгоритму [71]:

1) в качестве скрытого состояния q_1^*, в котором находился случайный процесс в первый момент времени генерации последовательности, принимается реализация дискретной случайной величины, имеющей распределение, заданное с помощью вектора начального распределения вероятностей Π;

2) в качестве скрытого состояния $q_t^*, t = \overline{2,T}$, в котором находился случайный процесс в момент времени t генерации последовательности, принимается реализация дискретной случайной величины x, имеющей распределение $p(q_t = x \mid q_{t-1} = q_{t-1}^*)$, которое соответствует вероятностям переходов из A при условии $q_{t-1} = q_{t-1}^*$;

3.а) в случае СММ с дискретным распределением наблюдений, наблюдение $\boldsymbol{o}_t, t = \overline{1,T}$, сгенерированное случайным процессом в момент времени t принимается равным реализации дискретной случайной величины $\boldsymbol{x}$, имеющей распределение $p(\boldsymbol{x} \mid q = q_t^*)$, которое соответствует вероятностям переходов из B при условии $q = q_t^*$;

3.б) в случае СММ с непрерывной плотностью распределения наблюдений, наблюдение $\boldsymbol{o}_t, t = \overline{1,T}$, сгенерированное случайным процессом в момент времени t принимается равным реализации многомерной непрерывной случайной величины $\boldsymbol{x}$, имеющей распределение, представленное смесью многомерных нормальных распределений $f(\boldsymbol{x}) = \sum_{m=1}^{M} \tau_{q_t^* m} g(\boldsymbol{x}; \mu_{q_t^* m}, \Sigma_{q_t^* m})$, которое соответствует условной плотности распределения наблюдений из B в скрытом состоянии q_t^*.

1.8.2 Моделирование неполных последовательностей

Назовём неполной или «дефектной» последовательностью такую последовательность O, в которой значение некоторых наблюдений не определено (т. е. имеются пропуски). При этом, как уже было упомянуто, наличие пропусков определяется некоторыми внешними факторами: то есть изучаемый процесс породил всю последовательность полностью без пропусков, но мы имеем дело с той же самой последовательностью, в которой по некоторым причинам значение отдельных наблюдений неизвестны. Обозначим пропуск символом $\varnothing$. Последовательность длиной T, порождённую случайным процессом, описываемым СММ с дискретной плотностью распределения можно описать следующим образом: $O = \left\{ o_t \in V^*, t = \overline{1,T} \right\}$ $V^* = V \cup \{\varnothing\}$. Последовательность длиной T, порождённую случайным процессом, описываемым СММ с непрерывной плотностью распределения, состоящую из Z-мерных наблюдений, и в которой могут иметься пропуски можно обозначить $O = \left\{ o_t \in R^*, t = \overline{1,T} \right\}$, $R^* = \mathbb{R}^Z \cup \{\varnothing\}$. В следующих разделах мы будем рассматривать именно такие последовательности.

Для моделирования G пропусков ($G \in \mathbb{N}, 1 \le G \le T-1$) в последовательности $O = \{\boldsymbol{o}_1, \boldsymbol{o}_2, ..., \boldsymbol{o}_T\}$ воспользуемся следующим алгоритмом:

пока не сгенерировано G пропусков:

1) выбирается индекс t как реализация случайной величины, имеющей дискретное равномерное распределение на множестве $\{t \mid t \in \mathbb{N}, t \le T, \boldsymbol{o}_t \ne \varnothing\}$;

2) генерируется пропуск в исходной последовательности O:

$\boldsymbol{o}_t := \varnothing$.

Выводы по первой главе и постановка задач

В данной главе приводятся результаты изучения современного состояния проблемы анализа неполных последовательностей, описываемых скрытыми марковскими моделями, а также основные положения теории скрытых марковских моделей. Современное состояние вопроса таково, что в теории СММ тема анализа неполных последовательностей затронута лишь частично. Авторами не освещены вопросы обучения СММ по неполным последовательностям, теоретически не обоснованы используемые подходы, не проведен сравнительный анализ их эффективности по отношению к другим методам, не выявлены их преимущества и недостатки. К тому же эти подходы тесно привязаны только к одной предметной области: распознаванию речи и движений по видеоряду. Другие области применения не рассмотрены.

Для восполнения пробела в теории СММ в области анализа неполных последовательностей, следует разработать методы анализа неполных последовательностей, описываемых скрытыми марковскими моделями. Для достижения данной цели следует решить следующие актуальные задачи. Разработать и исследовать методы:

- восстановления и декодирования неполных последовательностей, описываемых скрытыми марковскими моделями;

- распознавания неполных последовательностей, описываемых скрытыми марковскими моделями;

- обучения скрытой марковской модели по неполным последовательностям;

- распознавания неполных последовательностей, описываемых близкими скрытыми марковскими моделями, обученными на неполных последовательностях.

ГЛАВА 2 РАЗРАБОТКА МЕТОДА ДЕКОДИРОВАНИЯ И ВОССТАНОВЛЕНИЯ НЕПОЛНЫХ ПОСЛЕДОВАТЕЛЬНОСТЕЙ, ОПИСЫВАЕМЫХ СКРЫТЫМИ МАРКОВСКИМИ МОДЕЛЯМИ

В данной главе описан новый метод на основе модифицированного алгоритма Витерби, позволяющий осуществлять декодирование неполных последовательностей, описываемых скрытыми марковскими моделями. Научно обосновано его применение для восстановления неполных последовательностей, описываемых скрытыми марковскими моделями, а также проведен сравнительный анализ эффективности разработанного метода и других методов декодирования и восстановления неполных последовательностей, описываемых скрытыми марковскими моделями. Кроме того, в данной главе доказана возможность применения метода для решения практических задач: восстановления данных двигательной активности и декодирования наиболее вероятного пути абонента по транспортному графу на основе последовательности его регистраций в мобильной сети.

2.1 Разработка формулы вероятности эмиссии с помощью маргинализации

Определение неполной последовательности приведено в п 1.8.2. Для получения алгоритма анализа неполных последовательностей с помощью СММ необходимо, прежде всего, обратиться к формулам, по которым производится анализ полных последовательностей (глава 1).

Очевидно, что вычисление значений вероятностей эмиссий $b_i(\boldsymbol{o}_t)$, $i=\overline{1,N}$, $t=\overline{1,T}$, которые используются как в алгоритмах декодирования и распознавания последовательностей, так и алгоритме обучения СММ, невозможно, если $\boldsymbol{o}_t=\varnothing$, так как не определено конкретное наблюдаемое значение, а, следовательно, нельзя рассчитать значение $b_i(\boldsymbol{o}_t)$, которое соответствует данному наблюдению. Чтобы использовать эти формулы в случае неполных последовательностей, необходимо доопределить значение вероятности $b_i(\varnothing)$, $i=\overline{1,N}$ для тех вероятностей, которые рассчитываются по отсутствующим в последовательности наблюдениям.

Идея, лежащая в основе предлагаемых в данной работе методов состоит в принятии допущения, что на месте пропуска может стоять любое наблюдение из области допустимых значений наблюдений: дискретного алфавита V для СММ с дискретным распределением и R^Z для СММ с непрерывной плотностью распределения [72]. Руководствуясь этой идеей, представим $b_i(\varnothing),\ i=\overline{1,N}$ в виде маргинального распределение вероятности $b_i(x),\ i=\overline{1,N}$ при условии, что x неизвестен.

Для СММ с дискретным распределением представим значение $b_i(\varnothing),\ i=\overline{1,N}$ как сумму по всем возможным значениям пропущенного наблюдения:

$$b_i(\varnothing)=\sum_{v\in V} b_i(v)=1, \qquad i=\overline{1,N} \tag{59}$$

Аналогично для СММ с непрерывной плотностью распределения наблюдений представим значение $b_i(\varnothing),\ i=\overline{1,N}$ как интеграл (по всем Z измерениям) по всем возможным значениям пропущенного наблюдения:

$$b_i(\varnothing)=\int b_i(\boldsymbol{x})d\boldsymbol{x}=1, \qquad i=\overline{1,N} \tag{60}.$$

Руководствуясь теми же соображениями, определим значение плотности нормального распределения, входящего в смесь, для наблюдения-пропуска:

$$g(\varnothing,\mu_{im},\Sigma_{im})=\int g(\boldsymbol{x},\mu_{im},\Sigma_{im})d\boldsymbol{x}=1,\quad i=\overline{1,N},\ m=\overline{1,M}\,. \tag{61}$$

Справедливость формул (59)-(61) обусловлена тем, что в один момент времени имеется только одно наблюдение $\boldsymbol{x}$, а также тем, что $b_i(\boldsymbol{x})$ – вероятность символа $\boldsymbol{x}$ в дискретном случае или условная плотность распределения наблюдения $\boldsymbol{x}$ в непрерывном случае при нахождении в скрытом состоянии s_i, $i=\overline{1,N}$.

В сущности, сам способ вычисления функции правдоподобия для описываемой целой последовательности без пропусков, описываемой – это нахождение маргинального распределения наблюдений интегрированием функции правдоподобия по всем скрытым состояниям: $p(O\,|\,\lambda)=\sum_{q_1,q_2,\ldots,q_T} p\left(\{\boldsymbol{o}_1,\ldots,\boldsymbol{o}_T\},\{q_1,q_2,\ldots,q_T\}\,|\,\lambda\right)$. В данном случае пропусками считаются скрытые состояния.

2.2 Декодирование неполных последовательностей, описываемых скрытыми марковскими моделями, с помощью модифицированного алгоритма Витерби

Для декодирования последовательностей, порождённых процессами, описываемыми СММ, то есть формирования наиболее вероятной последовательности скрытых состояний наблюдаемой последовательности, используется алгоритм Витерби, описанный в 1 главе диссертации. С применением идеи маргинализации пропущенных наблюдений, описанной в разделе 2.1, алгоритм Витерби был модифицирован для декодирования неполных последовательностей [73].

Пусть необходимо найти (декодировать) наиболее вероятную последовательность скрытых состояний $Q = \{q_1, \ldots, q_T\}$ по наблюдаемой неполной последовательности $O^* = \{\boldsymbol{o}_1, \ldots, \boldsymbol{o}_T\}$. В таком случае модифицированный алгоритм Витерби, решающий данную задачу, состоит из следующих шагов:

1) инициализация:

$$\delta_1(i) = \begin{cases} \pi_i, & \boldsymbol{o}_1 = \varnothing \\ \pi_i b_i(\boldsymbol{o}_1), & \textit{иначе} \end{cases}, \quad i = \overline{1, N};$$

$$\psi_1(i) = 0, \quad i = \overline{1, N};$$

2) индукция:

$$\delta_t(j) = \begin{cases} \max\limits_{1 \le i \le N} \left[\delta_{t-1}(i) a_{ij} \right], & \boldsymbol{o}_t = \varnothing \\ \max\limits_{1 \le i \le N} \left[\delta_{t-1}(i) a_{ij} \right] b_j(\boldsymbol{o}_t), & \textit{иначе} \end{cases}, \quad j = \overline{1, N}, \quad t = \overline{2, T};$$

$$\psi_t(j) = \arg\max_{1 \le i \le N} \left[\delta_{t-1}(i) a_{ij} \right], \quad j = \overline{1, N}, \; t = \overline{2, T};$$

3) завершение:

$$q_T = \arg\max_{1 \le i \le N} \left[\delta_T(i) \right];$$

4) рекурсивное определение наиболее вероятной последовательности скрытых состояний:

$$q_t = \psi_{t+1}\left(q_{t+1} \right), \quad t = \overline{T-1, \; 1}.$$

После завершения алгоритма получим сформированную наиболее вероятную последовательность скрытых состояний: $Q = \left\{ q_1, \cdots, q_T \right\}$.

2.3 Восстановление неполных последовательностей с использованием модифицированного алгоритма Витерби

Алгоритм декодирования неполных последовательностей (п. 2.2) можно применить для восстановления последовательностей, содержащих пропуски. Пусть имеется СММ λ, а также сгенерированная соответствующим ей процессом последовательность O, в которой образовались случайным образом пропуски.

Для восстановления пропусков в последовательности O применим вначале к ней метод декодирования последовательностей с пропусками, описанный в п. 2.2. Так можно найти наиболее вероятную последовательность скрытых состояний $Q = \left\{ q_1, \cdots, q_T \right\}$. После декодирования можно восстановить каждый пропуск используя найденное скрытое состояние. Заместим пропуск наиболее вероятным наблюдением, соответствующим этому скрытому состоянию.

В случае СММ с дискретной плотностью распределения наблюдений, пропуск в момент t с найденным скрытым состоянием $q_t = s_{i^*}$ замещается наблюдением $\boldsymbol{o}_t = \arg\max_{\boldsymbol{x} \in V} b_{i^*}(\boldsymbol{x})$. Однако, для большей вариативности и приближенности восстановленных последовательностей к реальным, целесообразно заменять пропуск реализацией дискретной случайной величины, соответствующей i^* -му состоянию скрытой марковской модели, т.е. имеющей распределение $b_{i^*}(\boldsymbol{x})$.

В случае СММ с непрерывной плотностью распределения наблюдений пропуск в момент t с найденным скрытым состоянием $q_t = s_{i^*}$ замещается наблюдением $\boldsymbol{o}_t = \arg\max_{\boldsymbol{x} \in R^*} b_{i^*}(\boldsymbol{x}) = \arg\max_{\boldsymbol{x} \in R^*} \sum_{m=1}^{M} \tau_{i^* m} g(\boldsymbol{x}; \mu_{i^* m}, \Sigma_{i^* m})$. Очевидно, что в том случае,

когда распределение наблюдений описывается смесями нормальных распределений, максимум в данном выражении будет достигнут при $\boldsymbol{x} = \mu_{im^*}$, где $m^* = \arg\max_m (\tau_{im})$, т.е. наиболее вероятным значением $\boldsymbol{x}$ будет математическое ожидание компоненты смеси нормальных распределений, имеющей наибольший вес. К сожалению, такой выбор значения для замещения пропуска, в случае дальнейшего использования восстановленных последовательностей для обучения СММ приводит к сильному переобучению моделей, вызывающему вырождение ковариационных матриц, как показали проведённые автором эксперименты. Под переобучением в данном случае подразумевается явление из теории машинного обучения, когда обученная модель слишком хорошо объясняет обучающую выборку, что приводит к тому, что она плохо работает на новых примерах из тестовой выборки, которые она не встречала в момент обучения. Более целесообразно заменять пропуск реализацией непрерывной случайной величины, соответствующей i^* -му состоянию скрытой марковской модели, т.е. имеющей распределение $b_{i^*}(\boldsymbol{x})$ [74].

2.4 Восстановление неполных последовательностей с помощью значений соседних наблюдений

Примитивным способом восстановления пропущенных наблюдений является их замещение некоторым значением, полученным на основании значений соседних с пропуском непропущенных наблюдений. В случае СММ с дискретным распределением наблюдений пропуск можно замещать модой k соседних наблюдений. В случае СММ с непрерывным распределением наблюдений пропуск можно замещать средним арифметическим k соседних наблюдений. После восстановления последовательности таким способом некоторые пропуски могут остаться невосстановленными (к примеру, такие пропуски, у которых k соседних наблюдений тоже пропуски). Поэтому данная процедура выполняется повторно, но число рассматриваемых соседей k при этом увеличивается до размера всей последовательности T .

В данной диссертационной работе пропуск замещался на основании значений $k=10$ ближайших соседей (5 соседей слева и 5 справа). Данное число соседей было выбрано экспериментально: при таком значении параметра описанный выше алгоритм показал наилучшее качество восстановления.

2.5 Исследование модифицированного алгоритма Витерби при декодировании неполных последовательностей

2.5.1 Оценка эффективности модифицированного алгоритма Витерби при декодировании неполных последовательностей, описываемых скрытыми марковскими моделями с дискретным распределением наблюдений

В данном эксперименте проведена оценка эффективности алгоритмов декодирования последовательностей дискретных наблюдений, содержащих пропуски. В качестве истинной СММ была взята модель λ со следующими характеристиками. Число скрытых состояний $N=3$, размерность алфавита наблюдаемых символов $M=3$. Вектор распределения начального состояния: $\Pi=[1,0,0]$, матрица вероятностей переходов: $A=\begin{bmatrix}0.1 & 0.7 & 0.2\\ 0.2 & 0.2 & 0.6\\ 0.8 & 0.1 & 0.1\end{bmatrix}$, матрица эмиссии: $B=\begin{bmatrix}0.1 & 0.1 & 0.8\\ 0.1 & 0.8 & 0.1\\ 0.8 & 0.1 & 0.1\end{bmatrix}$.

С помощью модели λ было сгенерировано $K=100$ последовательностей наблюдений длиной $T=100$ с пропусками. Для декодирования использовалась истинная модель λ. Фиксировалось количество правильно декодированных скрытых состояний.

Рисунок 4 содержит результаты описанного выше эксперимента. Приведены усредненные значения после 100 запусков. Тип линии обозначает использованный метод декодирования: пунктирная – декодирование с помощью модифицированного алгоритма Витерби (п. 2.1), штрихпунктирная – восстановление пропусков по моде ближайших соседей (п. 2.4) и затем декодирования восстановленной последовательности с помощью стандартного алгоритма Витерби.

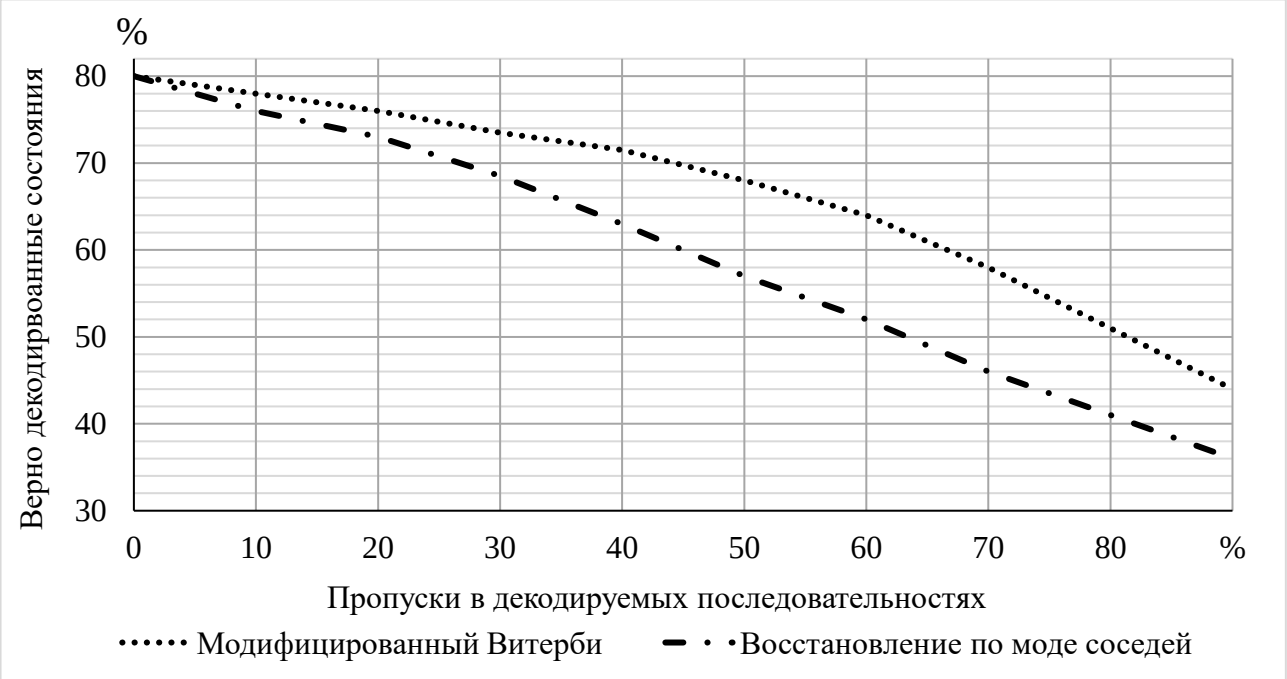

Рисунок 4 – Эффективность модифицированного алгоритма Витерби для декодирования неполных последовательностей дискретных наблюдений

Как видно из рисунка 4, метод декодирования с помощью модифицированного алгоритма Витерби позволяет до 1.3 раз увеличить количество верно декодированных скрытых состояний по сравнению со стандартным подходом, основанным на восстановлении пропусков по моде ближайших соседей, и последующим декодированием [73].

2.5.2 Оценка эффективности модифицированного алгоритма Витерби при декодировании неполных последовательностей, описываемых скрытыми марковскими модлеями с непрерывным распределением наблюдений

В вычислительном эксперименте проведено сравнение алгоритмов декодирования последовательностей векторов вещественных чисел, содержащих пропуски. В качестве истинной СММ была выбрана модель λ со следующими характеристиками. Число скрытых состояний $N = 3$, количество компонент в смесях

$M = 3$. Размерность векторов наблюдений $Z = 2$. Вектор распределения начального состояния: $\Pi = [1,0,0]$, матрица вероятностей переходов: $A = \begin{bmatrix} 0.1 & 0.7 & 0.2 \\ 0.2 & 0.2 & 0.6 \\ 0.8 & 0.1 & 0.1 \end{bmatrix}$, веса компонент смесей $\left\{\tau_{im}, i = \overline{1,N}, m = \overline{1,M}\right\} = \begin{pmatrix} 0.3 & 0.4 & 0.3 \\ 0.3 & 0.4 & 0.3 \\ 0.3 & 0.4 & 0.3 \end{pmatrix}$ (номеру строки соответствует номер скрытого состояния, а номеру столбца – номер компоненты смеси), вектора математических ожиданий компонент смесей $\left\{\mu_{im}, i = \overline{1,N}, m = \overline{1,M}\right\} = \begin{pmatrix} (0 \quad 0)^T & (1 \quad 1)^T & (2 \quad 2)^T \\ (3 \quad 3)^T & (4 \quad 4)^T & (5 \quad 5)^T \\ (6 \quad 6)^T & (7 \quad 7)^T & (8 \quad 8)^T \end{pmatrix}$ (номеру строки соответствует номер скрытого состояния, а номеру столбца – номер компоненты смеси), все ковариационные матрицы компонент смесей $\left\{\Sigma_{im}, i = \overline{1,N}, m = \overline{1,M}\right\}$ были выбраны единичными. С помощью модели λ из было сгенерировано $K = 100$ последовательностей наблюдений длиной $T = 100$ с пропусками. Для декодирования использовалась истинная модель λ. Фиксировался процент правильно декодированных скрытых состояний.

Рисунок 5 содержит результаты описанного выше эксперимента. Приведены средние значения после 100 запусков. Тип линии обозначает использованный метод декодирования: пунктирная – декодирование с помощью модифицированного алгоритма Витерби (п. 2.3), штрихпунктирная – восстановление пропусков по среднему арифметическому ближайших соседей (п. 2.4) и затем декодирование восстановленной последовательности с помощью стандартного алгоритма Витерби.

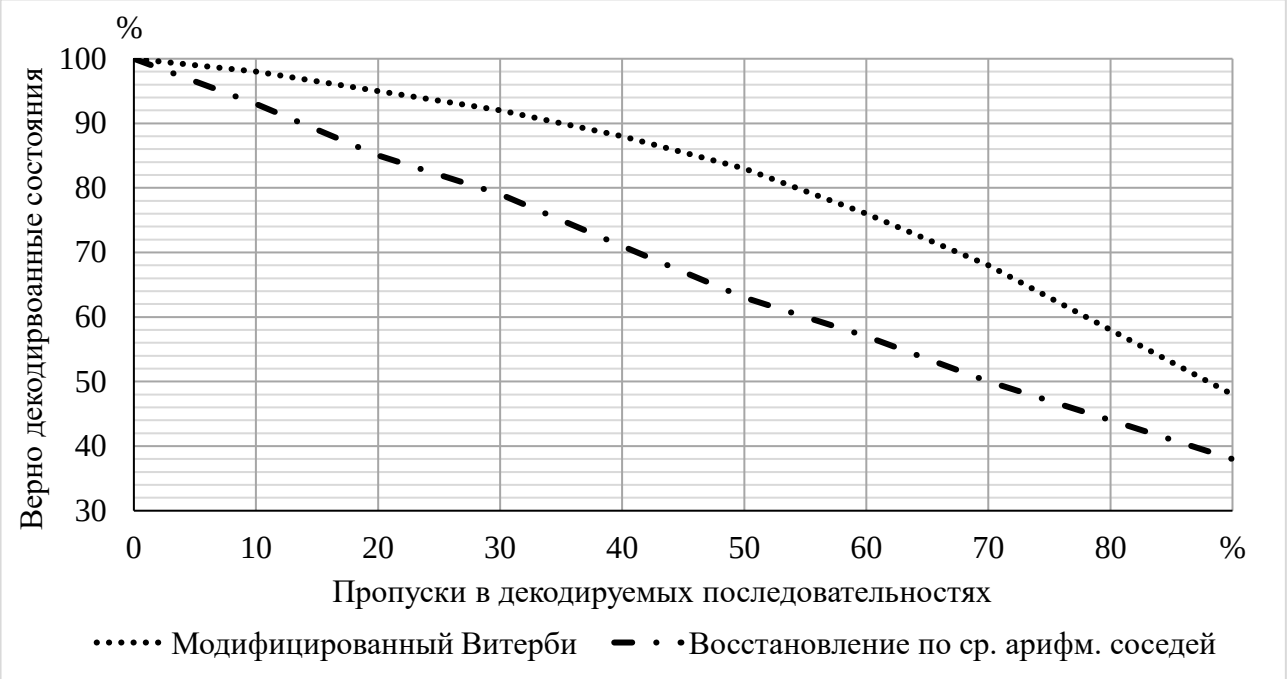

Рисунок 5 – Эффективность модифицированного алгоритма Витерби для декодирования неполных последовательностей векторов вещественных чисел

Из рисунка 5 видно, что метод декодирования с помощью модифицированного алгоритма Витерби позволяет до 1.4 раз увеличить количество правильно декодированных состояний по сравнению с подходом, основанным на восстановлении пропусков по среднему арифметическому ближайших соседей с последующим декодированием [75].

2.6 Исследование алгоритма восстановления неполных последовательностей, основанного на модифицированном алгоритме Витерби

2.6.1 Оценка эффективности алгоритма восстановления неполных последовательностей дискретных наблюдений, основанного на модифицированном алгоритме Витерби

Проведён эксперимент для оценки эффективности алгоритмов восстановления последовательностей дискретных наблюдений с пропусками. В качестве истинной СММ была выбрана модель λ со следующими характеристиками. Число скрытых состояний $N = 3$, размерность алфавита наблюдаемых символов $M = 3$. Вектор

распределения начального состояния: $\Pi = [1,0,0]$, матрица вероятностей перехо-

дов: $A = \begin{bmatrix} 0.1 & 0.7 & 0.2 \\ 0.2 & 0.2 & 0.6 \\ 0.8 & 0.1 & 0.1 \end{bmatrix}$, матрица эмиссии: $B = \begin{bmatrix} 0.1 & 0.1 & 0.8 \\ 0.1 & 0.8 & 0.1 \\ 0.8 & 0.1 & 0.1 \end{bmatrix}$. С помощью модели

λ было сгенерировано $K = 100$ последовательностей наблюдений длиной $T = 100$ с пропусками. Последовательности с пропусками генерировались таким же образом, как и в предыдущем эксперименте с помощью модели λ. Для восстановления использовалась истинная модель λ. Фиксировалось количество правильно восстановленных наблюдений.

Рисунок 6 содержит результаты описанного выше эксперимента. Приведены средние значения после 100 запусков. Тип линии обозначает использованный метод восстановления: пунктирная – восстановление с помощью модифицированного алгоритма Витерби (п. 2.3), штрихпунктирная – восстановление пропусков по моде ближайших соседей (п. 2.4).

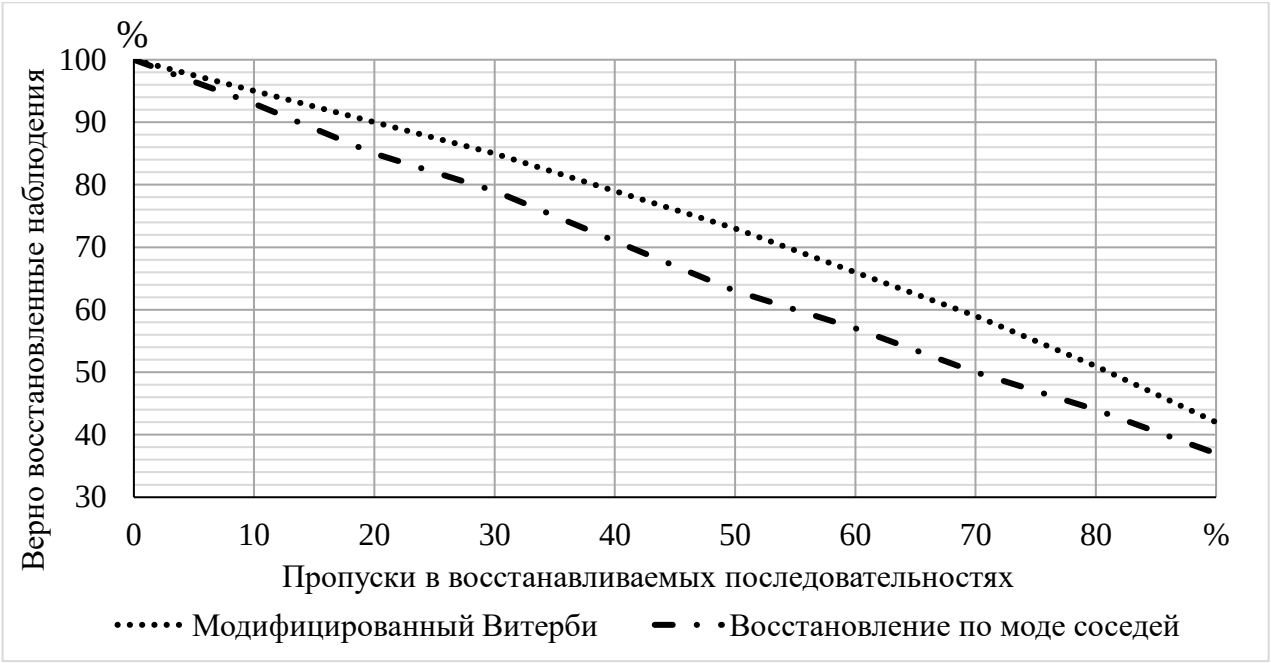

Рисунок 6 – Эффективность модифицированного алгоритма Витерби для восстановления неполных последовательностей дискретных наблюдений

В результате эксперимента, как видно на диаграмме зависимости на рисунке 6, доказано, что метод восстановления последовательностей с пропусками с

помощью модифицированного алгоритма Витерби позволяет до 1.2 раз увеличить количество верно восстановленных последовательностей по сравнению со стандартным подходом, основанным на восстановлении пропусков по моде ближайших соседей [73].

2.6.2 Оценка эффективности алгоритма восстановления неполных последовательностей векторов вещественных чисел, основанного на модифицированном алгоритме Витерби

Представлены результаты эксперимента по сравнению алгоритмов восстановления последовательностей векторов вещественных чисел, содержащих пропуски. В качестве истинной СММ была взята модель λ со следующими характеристиками. Число скрытых состояний $N = 3$, количество компонент в смесях $M = 3$. Размерность векторов наблюдений $Z = 2$. Вектор распределения начального состояния: $\Pi = [1,0,0]$, матрица вероятностей переходов: $A = \begin{bmatrix} 0.1 & 0.7 & 0.2 \\ 0.2 & 0.2 & 0.6 \\ 0.8 & 0.1 & 0.1 \end{bmatrix}$, веса компонент смесей $\left\{\tau_{im}, i = \overline{1,N}, m = \overline{1,M}\right\} = \begin{pmatrix} 0.3 & 0.4 & 0.3 \\ 0.3 & 0.4 & 0.3 \\ 0.3 & 0.4 & 0.3 \end{pmatrix}$ (номеру строки соответствует номер скрытого состояния, а номеру столбца – номер компоненты смеси), вектора математических ожиданий компонент смесей $\left\{\mu_{im}, i = \overline{1,N}, m = \overline{1,M}\right\} = \begin{pmatrix} (0 \quad 0)^T & (1 \quad 1)^T & (2 \quad 2)^T \\ (3 \quad 3)^T & (4 \quad 4)^T & (5 \quad 5)^T \\ (6 \quad 6)^T & (7 \quad 7)^T & (8 \quad 8)^T \end{pmatrix}$ (номеру строки соответствует номер скрытого состояния, а номеру столбца – номер компоненты смеси), все ковариационные матрицы компонент смесей $\left\{\Sigma_{im}, i = \overline{1,N}, m = \overline{1,M}\right\}$ были выбраны единичными. С помощью модели λ из было сгенерировано $K = 100$ последовательностей наблюдений длиной $T = 100$ с пропусками. Для восстановления была

использована истинная модель λ. Фиксировалась разница d между исходными и восстановленными наблюдениями, выраженная в средней арифметической норме разностей исходных и восстановленных векторов наблюдений:

$$d = \frac{\sum_{k=1}^{K} \left\| o_k - \hat{o}_k \right\|}{K}. \qquad (62)$$

Рисунок 7 содержит результаты описанного выше эксперимента. Приведены средние значения после 100 запусков. Тип линии обозначает использованный метод восстановления: пунктирная – восстановление с помощью модифицированного алгоритма Витерби (2.3), штрихпунктирная – восстановление пропусков по среднему арифметическому ближайших соседей (2.4).

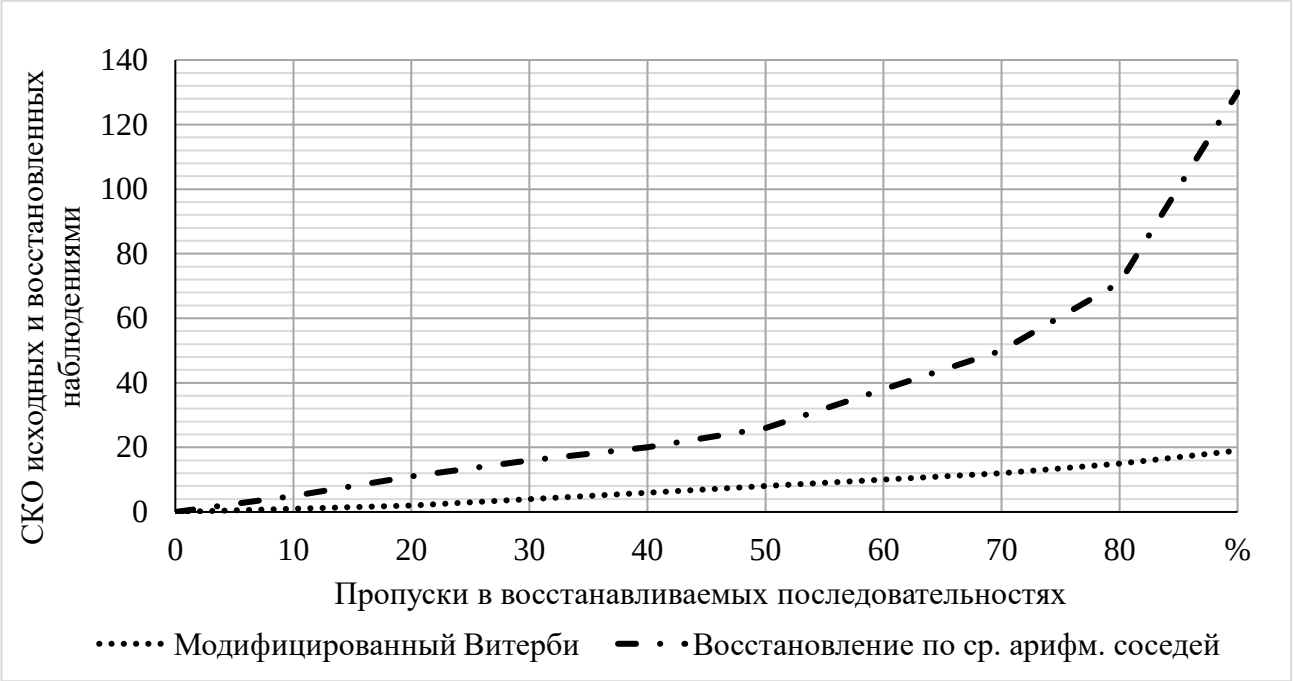

Рисунок 7 – Эффективность модифицированного алгоритма Витерби для восстановления неполных последовательностей векторов вещественных чисел

Как показано на диаграмме зависимости на рисунке 7, метод восстановления последовательностей с пропусками с помощью модифицированного алгоритма Витерби обеспечивает точность восстановления пропущенных наблюдений до 7 раз выше, чем подход, основанный на восстановлении пропусков по среднему арифметическому ближайших соседей [75].

2.7 Разработка методики восстановления неполных данных двигательной активности человека

Данная практическая задача заключается в восстановлении неполных последовательностей, генерируемых носимым устройством, анализирующим данные двигательной активности. Эта задача актуальна так как многие «умные» носимые устройства собирают различные данные о двигательной активности пользователя и для её корректной обработки нуждаются в высоком качестве данных и их целостности.

Для вычислительного эксперимента использовался набор данных “User Identification From Walking Activity” (распознавание пользователя по данным его/её двигательной активности), свободно доступный в интернете [29]. В наборе данных содержится информация, генерируемая смартфоном на базе операционной системы Android, расположенным в нагрудном кармане. Фиксировались показатели акселерометра телефона, в то время как каждый из участников эксперимента шёл по определённому заранее маршруту. Всего в эксперименте участвовало 22 человека. Рассматриваемый набор данных, по словам авторов, предназначен для исследований в области распознавания движений и идентификации людей на основе выявленных закономерностей их двигательной активности.

Данные для каждого участника эксперимента представлены в виде таблиц со следующими столбцами: «момент времени», «ускорение по оси x», «ускорение по оси y», «ускорение по оси z». Показания акселерометра были собраны с частотой 33 Гц. Таким образом, данные для отдельного пользователя могут быть представлены в виде последовательности трёхмерных векторов. Продолжительность сеанса измерений для отдельного пользователя достигает от 30 секунд до 11 минут. Рисунок 8 содержит визуализацию короткого отрезка данных акселерометра.

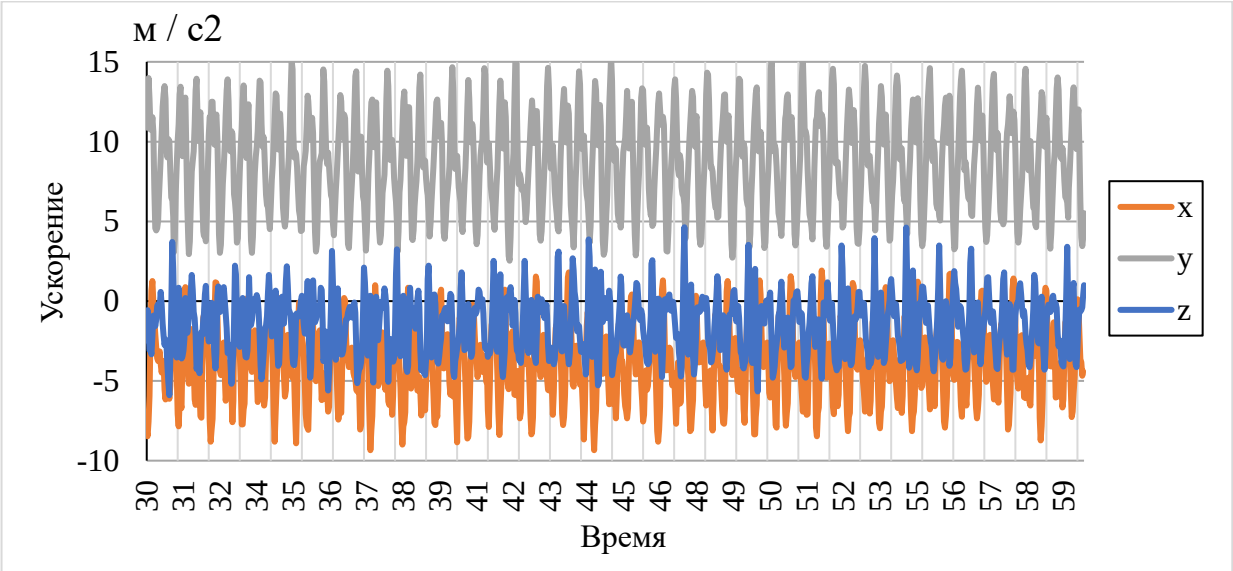

Рисунок 8 – Фрагмент данных акселерометра

Для каждого участника эксперимента была обучена своя СММ. Каждая СММ имела $N = 3$ скрытых состояния, $M = 3$ компонент смесей трёхмерных ($Z = 3$) нормальных распределений. Количество скрытых состояний и компонент смесей было подобрано эмпирически таким образом, чтобы достигалось максимальное качество восстановления последовательностей при приемлемом времени обучения СММ. Каждая последовательность трёхмерных векторов была разбита на подпоследовательности длиной $T = 100$ (что примерно соответствует трём секундам наблюдений). Для обучения СММ были взяты 75% случайно выбранных последовательностей из каждого класса, а оценивание качества алгоритма восстановления проводилось по оставшимся 25% последовательностей (тестовая выборка). В каждой тестовой последовательности определённый процент случайно выбранных наблюдений был заменён пропусками.

В качестве метрики качества восстановления взята среднеквадратичная норма разности между исходными и восстановленными векторами наблюдений, вычисляемая по формуле (62). Метрика вычислялась для различного количества пропущенных наблюдений в тестовой выборке и для двух различных методов восстановления пропущенных наблюдений: с помощью модифицированного алгоритма Витерби и с помощью среднего арифметического соседних с пропуском наблюдений. Рисунок 9 содержит результаты описанного выше эксперимента.

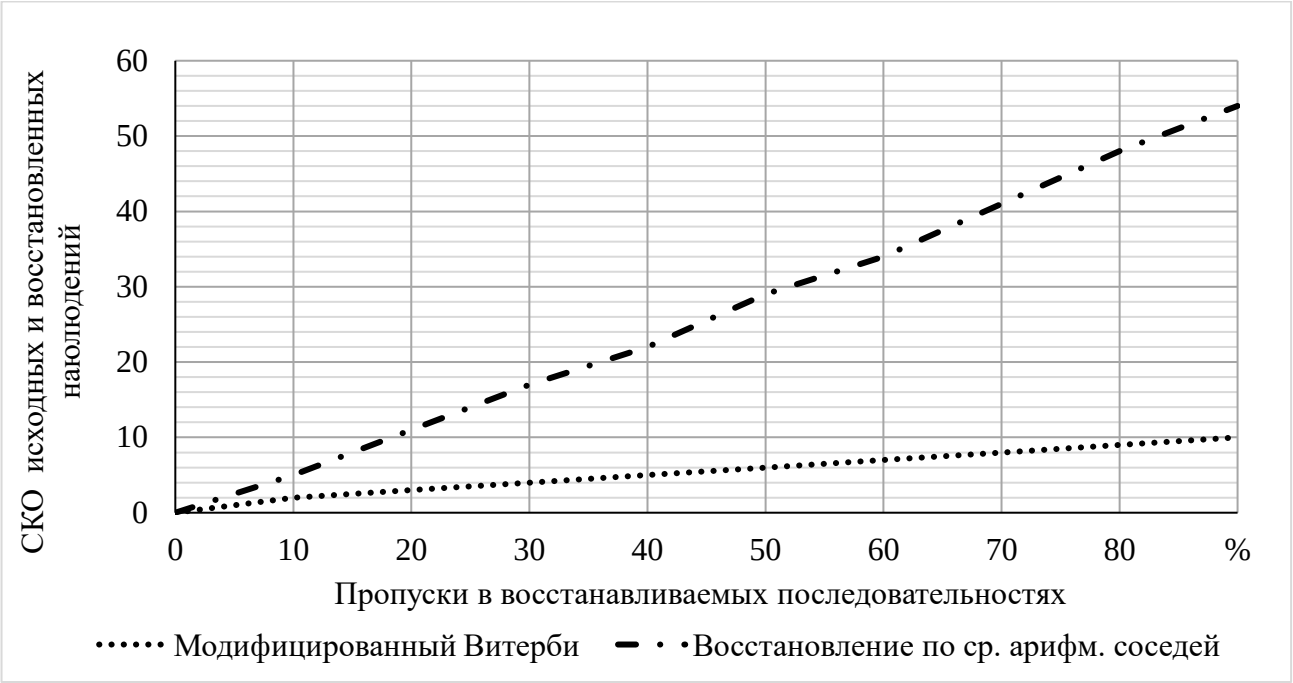

Рисунок 9 – Эффективность модифицированного алгоритма Витерби для восстановления неполных данных двигательной активности человека

Пунктирной линией обозначены результаты восстановления с помощью модифицированного алгоритма Витерби, а штрихпунктирной линией – с помощью среднего арифметического соседних с пропуском наблюдений [74].

Исходя из результатов данного эксперимента видно, что эффективность метода восстановления с использованием модифицированного алгоритма Витерби до 5 раз превосходит эффективность стандартного метода. Таким образом, основываясь на результатах экспериментов, можно рекомендовать использование модифицированного алгоритма Витерби для восстановления неполных последовательностей двигательной активности, описываемых СММ. Данный алгоритм превосходит стандартный метод, прост в реализации и не требует больших вычислительных затрат.

2.8 Разработка методики декодирования наиболее вероятного пути движения абонента по транспортному графу на основе последовательности регистраций в мобильной сети

Благодаря технологическому прогрессу на данный момент практически у каждого человека имеется персональное устройство сотовой связи, которое находится в непосредственной близости от него 24 часа в сутки. Даже самые простые мобильные телефоны генерируют массу информации во время сеансов связи и смены зоны местоположения (location update), что открывает для оператора сотовой связи широкие возможности по анализу перемещения своих абонентов. Актуальной становится задача превращения последовательности фрагментов информации о местоположении индивидуума в последовательность рёбер (путь) на заранее заданном графе, моделирующем транспортную сеть географической области. В англоязычной литературе такая задача носит название «map matching» (сопоставление с картой). Результат решения данной задачи имеет множество практических применений, включая, к примеру, выбор оптимального расположения рекламных конструкций для таргетирования рекламы согласно группам потребителей.

Существует несколько подходов к решению задачи map matching: геометрический, учитывающий только близость геометки к рёбрам графа, топологический, принимающий во внимание также связность рёбер графа и вероятностный, учитывающий, помимо вышеупомянутого, вероятности перемещений между элементами графа [76]. Часто в основе вероятностного подхода лежат скрытые марковские модели (СММ). Несмотря на большое количество публикаций по задаче map matching, в том числе использующих СММ, в большинстве из них в качестве исходных последовательностей сигналов используются данные GPS (global positioning system – глобальная система навигации), которые являются намного более частыми и точными, чем регистрации в мобильной сети [32]. Исключение составляет, например, работа, где СММ используются для анализа мобильных данных [77]. Тем не менее, ни в одной из ранее опубликованных работ не рассматривается случай, когда

между двумя последовательными регистрациями в мобильной сети не удается проложить путь с учётом заданного ограничения на длину пути, которое неизбежно приходится вводить при массовом анализе передвижения миллионов пользователей мобильной связи.

Таким образом, версия задачи map matching, основанная на CMM, использующая последовательности регистраций в мобильной сети, а также допускающая пропущенную информацию о пути по графу между последовательными регистрациями, является малоизученной. Данное исследование ставит целью восполнить этот пробел.

2.8.1 Задача и исходные данные

Для корректной постановки задачи необходимо в первую очередь привести краткие сведения о принципе устройства мобильной сети, которая генерирует анализируемые сигналы. Непосредственно оборудование сотовой связи устанавливается на базовые станции (БС), представляющие собой в случае outdoor (уличных) БС вышки с антеннами в верхней части. Рисунок 10 содержит пример типичной БС, расположенной в городе. Рисунок 11 содержит пример типичной БС, расположенной за городом.

Рисунок 10 – Пример типичной городской базовой станции сотовой связи

Рисунок 11 – Пример типичной загородной базовой станции сотовой связи

Базовая станция, как правило, имеет три антенны, угол между направлениями которых составляет 120 градусов, благодаря чему БС покрывает всю поверхность вокруг себя. Одна антенна покрывает один участок земли, называемый сектором или сетевым элементом. Базовые станции располагаются таким образом, что их секторы образуют структуру из примыкающих друг другу гексагонов, напоминающую пчелиные соты. Рисунок 12 изображает примерную схему организации мобильной сети. Здесь треугольниками показаны базовые станции, стрелками показаны азимуты антенн БС, секторами кругов показаны сектора покрытия, а пунктирными линиями – шестиугольники воображаемых сот. Однако стоит учитывать, что данная схема идеализированная, и в реальности покрытия секторов пересекаются, имеют неправильную форму и подвержены влиянию ландшафта, строений, растительности и имеют прочие недостатки. На практике полигоны покрытия секторов описываются полигонами, которые строят радиопланировщики исходя из мощности антенн и особенностей местности. Базовые станции объединяются в группы,

называемые location area (зона местоположения) для быстрого поиска абонента в сети.

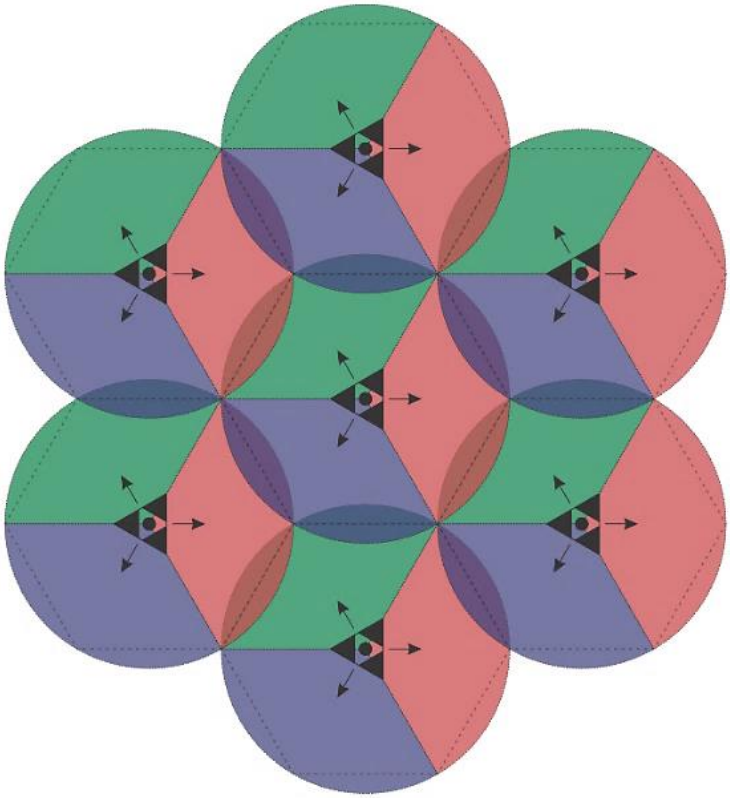

Рисунок 12 – Схема организации мобильной сети

Географическое положение мобильного устройства абонента можно определить во время следующих событий (условимся называть их регистрациями): входящего или исходящего голосового звонка, входящего или исходящего SMS-сообщения (SMS – short message service), во время генерации устройством интернет-трафика, а также во время процедуры location update (смены зоны местоположения) или сокращённо LU, которая происходит при включении устройства или переходе устройства из одной зоны местоположения в другую. Причем, сложность состоит в том, что фактически в момент события положение устройства можно определить только с точностью до сектора мобильной связи, в котором произошло событие.

Простым решением для восстановления истинного маршрута абонента является соединение центроид последовательных во времени регистраций абонента. Однако данное решение позволяет получить лишь очень грубый и приблизительный маршрут, который значительно отличается от реального маршрута абонента.

Гораздо более точно маршрут можно восстановить, сопоставив информацию о последовательных регистрациях абонентов с графом автомобильных дорог и пешеходных путей. Идея заключается в том, чтобы выбрать наиболее вероятную последовательность рёбер графа, исходя из последовательности регистраций, для каждой из которых известен полигон покрытия сектора и время регистрации. Данная задача, по сути, сводится к задаче, называемой map matching (сопоставление с картой). Рисунок 13 содержит пример траектории фактического перемещения абонента, а также покрытий секторов регистраций абонента во время этого перемещения и их соотношения с транспортным графом.

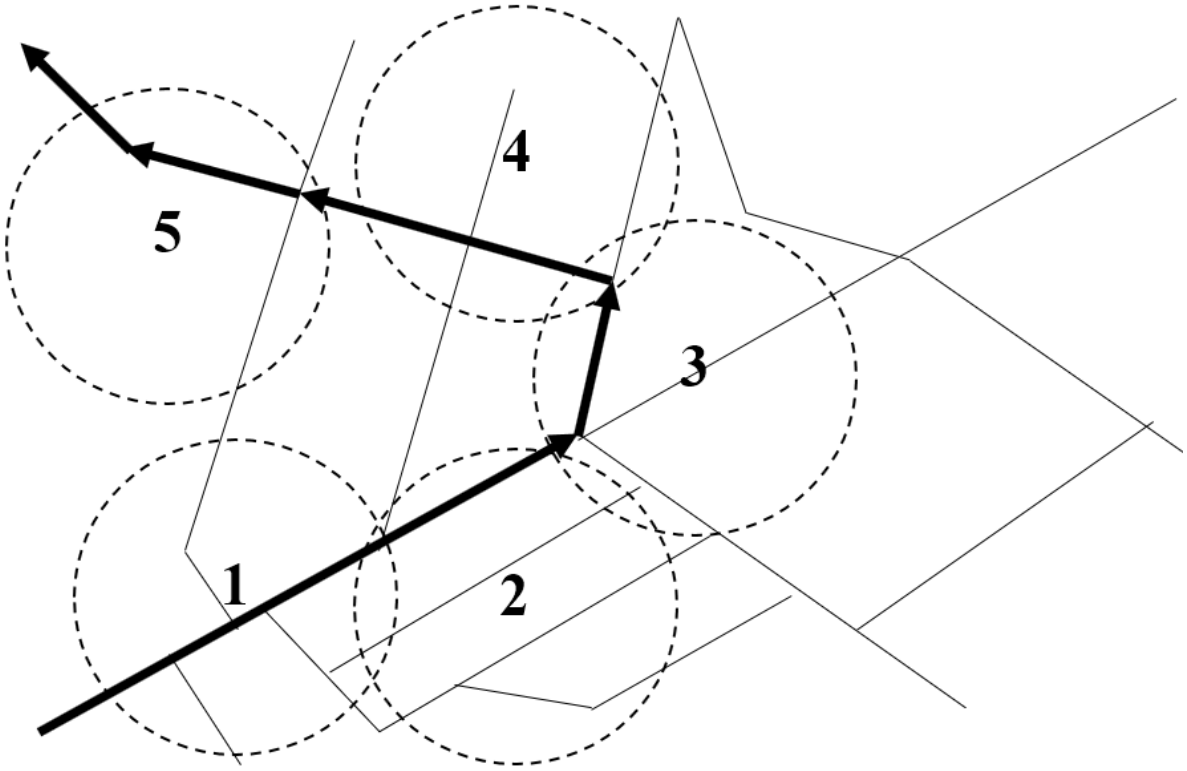

Рисунок 13 – Пример фрагмента транспортного графа с наложенными на него покрытиями секторов, на которых последовательно регистрировался абонент

Тонкими сплошными линиями показаны рёбра графа, жирными сплошными линиями показаны рёбра графа, по которым фактически двигался абонент, причём стрелка соответствует направлению движения, пунктирные окружности соответствуют покрытию сетевых элементов, на которых зарегистрировался абонент, а цифра в центре окружностей означает порядок регистрации во времени.

Таким образом, задача состоит в том, чтобы по последовательности регистраций абонента в мобильной сети выбрать наиболее вероятный маршрут его передвижения по транспортному графу. В данной работе используется транспортный граф, взятый из открытого картографического сервиса OSM (Open Street Maps) [78].

2.8.2 Алгоритм декодирования маршрута с использованием скрытых марковских моделей

Пусть дана последовательность регистраций мобильного устройства абонента на секторах мобильной сети $O = \left\{ o_t = \left(P_t, c_t \right) \middle| c_t \in R^2, P_t \subset R^2, t = \overline{t_1, t_T} \right\}$, где наблюдение $o_t,\ t = \overline{t_1, t_T}$ характеризуется временем регистрации t, полигоном покрытия сектора $P_t \subset R^2$, а также координатами базовой станции $c_t \in \mathbb{R}^2$.

Также пусть дан направленный граф $G = \left\{ V_G, E \right\}$, состоящий из набора вершин $V_G = \left\{ v_i \in R^2 \,|\, i = \overline{1, N_G} \right\}$ с известными географическими координатами, а также набора рёбер, соединяющих пары вершин $E_G = \left\{ e_k = \left\{ \left(v_i, v_j \right), d_k \right\} \middle| v_i, v_j \in V_G,\ w_k \in R,\ k = \overline{1, K_G} \right\}$, причём для каждого ребра известен его вес, который интерпретируется как его длина в километрах $w_k \in R$.

Необходимо найти такую последовательность вершин $\left\{ \hat{v}_{t_1}, \hat{v}_{t_2}, \ldots \hat{v}_{t_T} \right\}$, которая с наибольшей вероятностью соответствует последовательности регистраций O.

Далее приводится предлагаемый алгоритм решения данной практической задачи с помощью скрытых марковских моделей:

1) для каждого наблюдения $o_t, t = \overline{t_1, t_T}$ найдём его вероятность при условии нахождения абонента в момент времени t на одной из вершин графа $p\left(o_t \,|\, v_i \right),\ i = \overline{1, N_G}, t = \overline{t_1, t_T}$. Для расчёта вероятности будем использовать отдалённость вершины от координат базовой станции c_t, считая её наиболее вероятным местом нахождения абонента, при этом для вершин, не попадающих в область полигона, будем считать данную вероятность равной нулю:

$$p(o_t \mid v_i) = \begin{cases} \dfrac{1}{\sqrt{2\pi}\sigma} e^{-0.5\left(\frac{\|c_t - v_i\|_{ортодромии}}{\sigma}\right)^2}, & v_i \in P_t, \quad i = \overline{1, N_G}, \\ 0, & v_i \notin P_t \end{cases}$$

где оператор $\| \|_{ортодромии}$ означает длину ортодромии (кратчайшего расстояния) между двумя точками на земной поверхности, и где присутствует неизвестный параметр σ, который можно интерпретировать как среднее расстояние между центроидой покрытия сектора, на котором зарегистрировался абонент, и его фактическим местом нахождения в данный момент. Процедура оценивания параметра σ будет рассмотрена в следующем разделе;

2) для каждой пары последовательных наблюдений $\left(o_{t_\tau}, o_{t_{\tau+1}}\right), \tau = \overline{1,T}$ рассчитаем вероятность перехода абонента из одной из вершин графа v_i, попадающей в полигон покрытия в момент времени t_τ, в любую другую вершину графа (в том числе ту же самую) v_j, попадающую в полигон покрытия в момент времени $t_{\tau+1}$:

$$p_{t_\tau}\left(v_i, v_j\right) = \begin{cases} \dfrac{1}{\beta} e^{\frac{-\left|w_{v_i v_j}\right|}{\beta}}, & v_i \in P_{t_t} \; u \; v_j \in P_{t_{\tau+1}} \; u \; \exists \, w_{v_i v_j} \in W \\ 0, & v_i \notin P_{t_\tau} \; или \; v_j \notin P_{t_{\tau+1}} \; или \; \neg \exists w_{v_i v_j} \in W \end{cases}, \quad \begin{matrix} i, j = \overline{1, N_G} \\ \tau = \overline{1,T} \end{matrix},$$

что соответствует экспоненциальному распределению, где $w_{v_i v_j}$ – длина кратчайшего пути от вершины v_i до вершины v_j транспортного графа G, а W - множество всех кратчайших путей между парами вершин, которые возможно построить на графе G. Если пути между вершинами не удаётся найти, то вероятность перехода также считается нулевой. Процедура оценивания параметра экспоненциального распределения β будет рассмотрена далее.

Поскольку разрабатываемый алгоритм предполагается использовать для декодирования маршрутов миллионов абонентов на реальном транспортном графе, содержащим огромное количество вершин (для оценки масштабов, OSM граф дорог города Москвы в пределах МКАД содержит около миллиона вершин), поиск

пути между вершинами должен производиться с учётом определённых ограничений на максимальную возможную длину пути. В данной работе предлагается использовать ограничение, пропорциональное расстоянию между координатами последовательных базовых станций. Таким образом, между вершинами, соответствующими каждой паре последовательных наблюдений $\left(o_{t_\tau}, o_{t_{\tau+1}}\right), \tau = \overline{1,T}$ будет производится поиск путей длиной не более $C \cdot \left\| c_{t_\tau} - c_{t_{\tau+1}} \right\|_{ортодромии}$, где $C \geq 1$ - некоторая константа, которую можно интерпретировать как множитель, позволяющий учесть тот факт, что путь по графу как правило длиннее прямого расстояния между базовыми станциями.

Вводимое таким образом ограничение неизбежно приводит к ситуациям, когда между вершинами, принадлежащими некоторым парам последовательных наблюдений, не получается построить ни одного пути по графу. Существует два выхода из данной ситуации. В первом случае можно отказаться от восстановления маршрута между данной парой регистраций и смириться с получившимся разрывом в маршруте, а во втором случае каким-либо образом попытаться восполнить данный пропуск и проложить наиболее правдоподобный маршрут. Особенность данной работы заключается в том, что используется именно второй вариант борьбы с пропусками. Идея заключается в том, чтобы присвоить переходам между каждой парой вершин некоторую константную вероятность (например, единицу). Таким образом, при дальнейшем декодировании наиболее вероятной последовательности данный переход не будет отдавать предпочтение определённой паре вершин, но при этом будет выбрана такая пара вершин, которая наиболее соответствует общему пути. Введём формулу вычисления модифицированных вероятностей переходов $p^*_{t_\tau}\left(v_i, v_j\right)$, учитывающих рассмотренную выше ситуацию:

$$p^*_{t_\tau}\left(v_i, v_j\right) = \begin{cases} p_{t_\tau}\left(v_i, v_j\right), \ \exists i^* \exists j^* : \exists w_{v_{i^*} v_{j^*}} \in W \\ 1, \ \neg\left(\exists i^* \exists j^* : \exists w_{v_{i^*} v_{j^*}} \in W\right) u \ v_i \in P_{t_\tau} \ u \ v_j \in P_{t_{\tau+1}} \\ 0, \ \neg\left(\exists i^* \exists j^* : \exists w_{v_{i^*} v_{j^*}} \in W\right) u \ (v_i \notin P_{t_\tau} \ или \ v_j \notin P_{t_{\tau+1}}) \end{cases}, \quad \begin{matrix} i, j = \overline{1, N_G} \\ \tau = \overline{1,T} \end{matrix}$$

3) для восстановления наиболее вероятной последовательности рёбер графа применим концепцию скрытых марковских моделей. В качестве скрытых состояний будут выступать вершины графа $v_i, i=\overline{1, N_G}$, в качестве наблюдений – регистрации абонента $o_t, t=\overline{t_1, t_T}$, в качестве распределения наблюдений $b_i(o_t)$ вероятности $p(o_t \mid v_i),\ i=\overline{1, N_G}, t=\overline{t_1, t_T}$, в качестве распределения начального скрытого состояния $\pi_i = p(q_1 = v_i)$ вероятность $p(o_1 \mid v_i),\ i=\overline{1, N_G}$, а в качестве вероятностей переходов из одного скрытого состояния в другое a_{tij} модифицированные вероятности переходов $p^*{}_t(v_i, v_j),\ i, j=\overline{1, N_G},\ t=\overline{t_2, t_T}$.

4) благодаря построению такого соответствия между элементами задачи и элементами СММ, как в пункте 3 алгоритма, становится возможным проводить декодирование последовательностей с пропущенными переходами с помощью модифицированного алгоритма Витерби, основанного на стандартном алгоритме Витерби [61], в результате применения которого можно получить наиболее вероятную последовательность скрытых состояний (вершин графа в терминологии задачи): $\{\hat{v}_{t_1}, \hat{v}_{t_2}, \ldots \hat{v}_{t_T}\}$.

Модифицированный алгоритм Витерби (для упрощения записей вместо индекса t_τ будем использовать индекс $t=\tau$, который можно интерпретировать как порядковый номер наблюдения):

пусть необходимо найти (декодировать) наиболее вероятную последовательность скрытых состояний $\hat{Q}_v = \{\hat{v}_1, \ldots, \hat{v}_T\}$ по наблюдаемой неполной последовательности $O^* = \{o_1, \ldots, o_T\}$. В таком случае модифицированный алгоритм Витерби, решающий данную задачу, состоит из следующих шагов:

а) инициализация:

$$\delta_1(i) = \pi_i b_i(\boldsymbol{o}_1),\quad i=\overline{1, N};$$

$$\psi_1(i) = 0,\quad i=\overline{1, N};$$

б) индукция:

$$\delta_t(j) = \max_{1 \le i \le N} \left[\delta_{t-1}(i) a_{tij} \right] b_j(\boldsymbol{o}_t), \quad j = \overline{1, N}, \quad t = \overline{2, T};$$

$$\psi_t(j) = \arg\max_{1 \le i \le N} \left[\delta_{t-1}(i) a_{tij} \right], \quad j = \overline{1, N}, \; t = \overline{2, T};$$

в) завершение:

$$\hat{v}_T = \arg\max_{1 \le i \le N} \left[\delta_T(i) \right];$$

г) рекурсивное определение наиболее вероятной последовательности скрытых состояний:

$$q_t = \psi_{t+1}\left(q_{t+1} \right), \quad t = \overline{T-1, \; 1}.$$

После завершения алгоритма получим сформированную наиболее вероятную последовательность скрытых состояний: $\hat{Q}_v = \left\{ \hat{v}_1, \cdots, \hat{v}_T \right\}$. Следует обратить внимание, что отличие модифицированного алгоритма Витерби от оригинального заключается в том, как вычисляются вероятности переходов a_{tij} в случаях, когда не удаётся найти ни одного пути по графу между вершинами в момент $t-1$ и t.

5) после декодирования оптимальной последовательности вершин следует достроить оптимальный путь по графу между последовательными вершинами $\hat{v}_{t_1}$ и $\hat{v}_{t_T}$ в виде промежуточных вершин с неизвестным временем. Данный путь вычисляется на графе без каких-либо ограничений его длины. Если даже в таком случае не удаётся найти ни одного пути между парой вершин, то маршрут разбивается на два отдельных. Неизвестное время у каждой из промежуточных вершин можно восстановить по известному фактическому времени между регистрациями пропорционально весу рёбер в пути, ведущем к промежуточной вершинам. Таким образом, в итоге можно восстановить полный путь абонента по графу $\left\{ \hat{v}_{t_1}, \hat{v}_{t_{1,1}}, \hat{v}_{t_{1,2}}, \ldots \hat{v}_{t_2}, \hat{v}_{t_{2,1}}, \hat{v}_{t_{2,2}} \ldots \hat{v}_{t_T} \right\}$.

2.8.3 Эталонные данные и оценивание неизвестных параметров

В качестве фактической траектории движения предлагается использовать анонимизированные данные, собранные с помощью специального приложения, устанавливаемого на телефоны, которое в момент регистрации устройства на базовой станции позволяет установить имя сектора, на котором произошла регистрация, время регистрации и GPS-координаты устройства в момент регистрации. Получаемую с помощью такой процедуру последовательность данных можно формализовать следующим образом:

$$\tilde{O} = \left\{ \tilde{o}_t = \left(\tilde{P}_t, \tilde{c}_t, z_t \right) \,\middle|\, \tilde{P}_t \subset R^2, \tilde{c}_t, z_t \in R^2, t = \overline{\tilde{t}_1, \tilde{t}_{\tilde{T}}} \right\},$$

где в каждый момент времени к данным регистрации, описанным выше, добавляется GPS-координата z_t устройства абонента в этот момент.

С помощью эталонных данных становится возможным оценить неизвестные параметры распределений σ и β, присутствующие в алгоритме, описанном в предыдущем пункте:

1) для оценивания параметра σ воспользуемся робастной оценкой MAD (median absolute deviation – медианное абсолютное отклонение) неизвестного стандартного отклонения нормального распределения [79]:

$$\sigma = 1.4826 \underset{t=\overline{\tilde{t}_1, \tilde{t}_{\tilde{T}}}}{\textit{медиана}} \left(\left\| \tilde{c}_t - z_t \right\|_{\textit{ортодромии}} \right); \tag{63}$$

2) для оценивания параметра β воспользуемся робастной оценкой параметра экспоненциального распределения, предложенной в [80]:

$$\beta = \frac{1}{\ln(2)} \underset{\tau=\overline{1,\tilde{T}}}{\textit{медиана}} \left(\left| w_{\tilde{v}_\tau \tilde{v}_{\tau+1}} \right| \right), \tag{64}$$

где $\tilde{v}_\tau \in V_G$, $\tau = \overline{1,\tilde{T}}$ - наиболее вероятная вершина графа в момент времени t_τ при условии наблюдения последовательности $\tilde{O}$. Последовательность наиболее вероятных вершин можно получить с помощью алгоритма map matching, описанного в предыдущих разделах за исключением того, что вместо полигона покрытия P_t и центроиды c_t всегда использовать GPS-координаты z_t.

Альтернативным способом подбора оптимальных параметров распределений σ и β является подбор данных параметров по сетке до тех пор, пока не будет достигнута максимальная метрика качества, алгоритм расчёта которой будет приведён в следующем разделе.

2.8.4 Алгоритм расчёта метрики качества решения задачи map matching

Для оценивания качества решения задачи map matching необходимо разработать метрику, которая учитывает, как пространственное, так и временное различие между фактической траекторией движения абонента и восстановленным полным путём по транспортному графу. За основу предлагаемого в данной работе алгоритме расчёта метрики было взято расстояние Фреше [81], которое позволяет оценить меру сходства двух кривых. Интуитивно расстояние Фреше между двумя кривыми можно интерпретировать как минимальную длину поводка, которая необходима для того, чтобы хозяин собаки, двигающийся по одной кривой, и собака, двигающаяся по другой кривой, смогли пройти свои кривые от начала до конца. В предлагаемой метрике алгоритм расчёта расстояния Фреше был значительно модифицирован, чтобы учитывать специфику задачи и вышеобозначенные требования к метрике.

Пусть дана эталонная траектория движения $\tilde{O}$ абонента и восстановленный путь движения абонента по графу $\hat{Q}_v = \left\{ \hat{v}_{t_1}, \hat{v}_{t_{1,1}}, \hat{v}_{t_{1,2}}, \ldots \hat{v}_{t_2}, \hat{v}_{t_{2,1}}, \hat{v}_{t_{2,2}} \ldots \hat{v}_{t_T} \right\}$. Необходимо найти среднее отклонение между эталоном и восстановленной последовательностью из наблюдаемых в одинаковый момент времени d, а также процент точек образца, для которых не нашлось соответствующего по времени отрезка точек восстановленного трека J_p.

Приведём далее алгоритм вычисления данного отклонения:

0) Инициализируем счётчик количества точек образца, для которых не нашлось соответствующего по времени отрезка точек восстановленного трека: $J = 0$;

1) для каждого элемента $z_t \in \tilde{O}$ и $\hat{v}_t \in \hat{V}$:

1.1) найдём пару соответствующих ему по времени элементов из другой последовательности, удовлетворяющих условию $\hat{v}_{t_\tau} \le z_t \le \hat{v}_{t_{\tau+1}}$ для z_t, и $z_{\tilde{t}_\tau} \le \hat{v}_t \le z_{\tilde{t}_{\tau+1}}$ для $\hat{v}_t$;

1.2) если пара элементов в пункте 1.1 была найдена,

то найдём точку, соответствующую времени t на другой последовательности:

$\hat{v}^* = \hat{v}_{t_\tau} + \left(t - t_\tau\right)\frac{\hat{v}_{t_{\tau+1}} - \hat{v}_{t_\tau}}{t_{\tau+1} - t_\tau}$ для z_t или $z^* = z_{\tilde{t}_\tau} + \left(t - \tilde{t}_\tau\right)\frac{z_{\tilde{t}_{\tau+1}} - z_{\tilde{t}_\tau}}{\tilde{t}_{\tau+1} - \tilde{t}_\tau}$ для $\hat{v}_t$;

если пара наблюдений в п. 1.1 не была найдена для точки образца z_t , то увеличим на единицу счётчик: $J = J + 1$

1.3) вычислим отклонение, равное $d\left(z_t\right) = \left\|\hat{v}^* - z_t\right\|_{ортодромии}$ для z_t или $d\left(\hat{v}_t\right) = \left\|z^* - \hat{v}_t\right\|_{ортодромии}$ для $\hat{v}_t$;

2) найдём медианное отклонение $d = \underset{e \in \tilde{O} \cup \hat{Q}_v}{median}\, d\left(e\right)$.

3) получим процент точек трека-образца, для которых не нашлось соответствующей точки на восстановленном треке: $J_p = 100\,\frac{J}{\tilde{T}}\%$.

Далее, при наличии множества траекторий $\tilde{O}^k, k = \overline{1,K}$ и множества соответствующих им отклонений d^k и процентов пропусков J_p^k, можно агрегировать их в единую метрику, например, взяв от них медиану: $\bar{d} = \underset{k}{median}\, d^k$, $\bar{J}_p = \underset{k}{median}\, J_p^k$.

2.8.5 Вычислительный эксперимент

Эксперимент проводился для 20000 анонимизированных суточных треков, которые содержали регистрации на базовых станциях, находящихся в пределах МКАД г. Москвы, и для которых имелись соответствующие GPS-треки образцы.

В первом эксперименте проводится сравнение двух версий алгоритма декодирования маршрута: в первом случае при отсутствии хотя бы одного пути по

графу между двумя последовательными регистрациями в треке, такой трек разбивается на два трека, каждый из которых декодировался бы отдельно с помощью стандартного алгоритма Витерби, а во втором случае используем модифицированный алгоритм Витерби для устранения таких разрывов в маршрутах. Таблица 1 содержит измеренное количество точек трека-образца, для которых нашлась соответствующая точка на восстановленном треке $\bar{J}_p$ в обоих случаях.

Таблица 1 – Оценка эффективности стандартной и модифицированной версий алгоритма Витерби для декодирования маршрута

Используемая метрика	Алгоритм Витерби	
	модифицированный	стандартный
Количество точек трека-образца, для которых нашлась соответствующая точка на восстановленном треке, %	82.5	51.1

Как видно из таблицы 1, модифицированный алгоритм Витерби, способный обрабатывать неизвестные вероятности переходов позволяет увеличить покрытие точек образца более чем на 30%, при этом сохранив ограничение на максимальную длину пути по графу.

Во втором эксперименте приводятся результаты подбора оптимальных оценок параметров распределений σ и β, используемых в алгоритме декодирования маршрутов с помощью модифицированного алгоритма Витерби. Значение параметра σ варьировалось от 0.5 до 3, а значение параметра β - от 0.5 до 2, при этом для каждой комбинации параметров вычислялась метрика качества $\bar{d}$ - медианное отклонение между декодированным треками и треками-образцами в метрах. Рисунок 14 содержит результаты данного эксперимента.

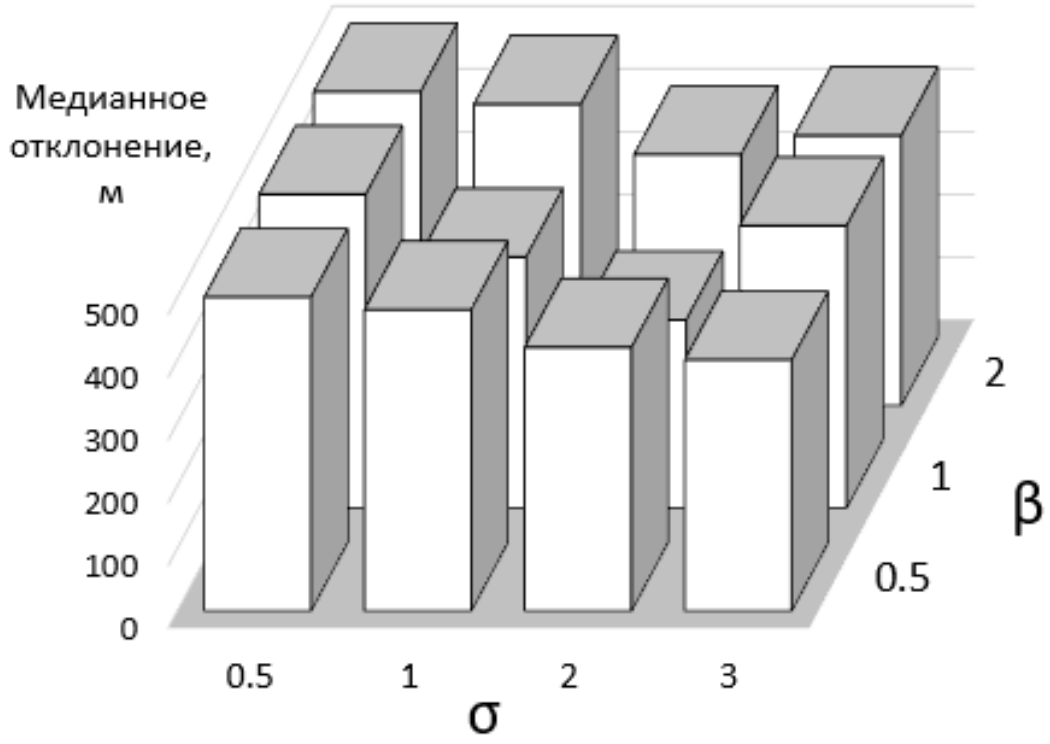

Рисунок 14 – Выбор оптимальных параметров распределений в алгоритме декодирования маршрутов

Как видно из рисунка 14, оптимальной комбинацией параметров является $\sigma = 2$, $\beta = 1$. Оптимальность найденной оценки параметра σ была также подтверждена с помощью формулы (63): согласно ней была получена оценка параметра $\sigma = 2.12$. В то же время вычисление параметра β по формуле (64) произведено не было, так для этого необходимо проводить процедуру map matching уже на треках-образцах, что требует в свою очередь подбора параметров распределений, а это невозможно сделать без более точного образца.

В третьем эксперименте проводится сравнение алгоритма декодирования оптимального маршрута, использующего скрытые марковские модели, и простого алгоритма, соединяющего координаты базовых станций в том порядке, в котором они встречаются в треке. Для каждого из исходных треков был получен наиболее вероятный маршрут с помощью алгоритма декодирования маршрута, основанного на СММ, а также с помощью простейшего последовательного соединения координат базовых станций между собой. Качество полученных маршрутов в обоих случаях

оценивалось с помощью медианного отклонения $\overline{d}$ между декодированным треками и треками-образцами в метрах. Таблица 2 содержит результаты данного эксперимента.

Таблица 2 – Эффективность новой методики декодирования (алгоритм на основе СММ) наиболее вероятного пути движения абонента по транспортному графу

Используемая метрика	Алгоритм	
	модифицированный Витерби	соединяющий координаты БС
Медианное отклонение между декодированными треками и треками-образцами, м	279	732

Как видно из таблицы 2, предложенная автором методика с вероятностным подходом на основе СММ, позволяет увеличить точность декодирования более чем в 2.5 раза по сравнению с простым алгоритмом, соединяющим координаты БС, и не учитывающим транспортный граф.

Таким образом, на основании теоретических выкладок разработана методика декодирования наиболее вероятного пути движения абонента по транспортному графу на основе последовательности регистраций в мобильной сети. Особенность данной методики заключается в том, что она позволяет обрабатывать данные регистраций в мобильной сети, а не GPS, а также устойчива к неполноте данных. Было показано, что её использование для решения данной задачи позволяет достичь увеличения точности декодирования более чем в 2.5 раза, по сравнению с простым алгоритмом соединения координат базовых станций последовательных регистраций абонента. соединения координат базовых станций последовательных регистраций абонента. Для решения поставленной задачи использован модифицированный алгоритм Витерби и доказана возможность находить с его помощью маршрут даже для тех участков, где не удаётся построить путь по графу с учётом заданных ограничений. Методика позволяет оптимизировать вычисления, при этом существенно увеличив процент покрытия суточного передвижения абонента по сравнению со стандартным алгоритмом Витерби (более чем на 30%) [82].

Выводы по второй главе

В данной главе автором предложен и научно обоснован новый метод декодирования и восстановления последовательностей с пропусками, основанный на модификации алгоритма Витерби для случая пропущенных наблюдений. Преимущество предложенного метода по сравнению с ранее известными подходами было подтверждено экспериментально. Стандартный подход (т. е. восстановление по моде в случае дискретных наблюдений или восстановление по среднему арифметическому соседей в случае наблюдений, являющихся векторами вещественных чисел), оказался менее эффективным в сравнении с новым.

Разработанный метод восстановления неполных последовательностей был применён для решения практической задачи восстановления неполных данных двигательной активности. Было показано, что его использование для решения данной задачи обеспечивает большую точность, чем замещение пропусков средним арифметическим соседних наблюдений (точность в 5 раз выше при 50% пропусков в последовательностях).

Модифицированный метод декодирования неполных последовательностей был применён для решения практической задачи декодирования наиболее вероятного пути движения абонента по транспортному графу на основе последовательности регистраций в мобильной сети. Доказано, что его использование для решения данной задачи обеспечивает в 2.5 раза большую точность, чем соединение центроид покрытий секторов последовательных регистраций абонента.

ГЛАВА 3 РАЗРАБОТКА МЕТОДА РАСПОЗНАВАНИЯ НЕПОЛНЫХ ПОСЛЕДОВАТЕЛЬНОСТЕЙ, ОПИСЫВАЕМЫХ СКРЫТЫМИ МАРКОВСКИМИ МОДЕЛЯМИ

В данной главе описан новый метод на основе модифицированного алгоритма forward-backward, который позволяет вычислять значение логарифма функции правдоподобия того, что случайный процесс, описываемый скрытой марковской моделью, сгенерировал неполную последовательность. Научно обосновано его применение для распознавания неполных последовательностей, описываемых скрытой марковской моделью. Проведен сравнительный анализ эффективности разработанного метода по отношению к другим методам распознавания неполных последовательностей, описываемых скрытыми марковскими моделями, в том числе и к представленному во второй главе методу на основе модифицированного алгоритма Витерби.

3.1 Распознавание неполных последовательностей с помощью модифицированного алгоритма forward-backward

Задача распознавания в данном случае имеет ту же постановку, что и в п. 1.4 за исключением того, что распознаваемая последовательность является неполной. Воспользуемся формулой эмиссии в случае пропущенного наблюдения (полученной в п. 2.1), чтобы доопределить алгоритм вычисления прямых и обратных вероятностей forward-backward до случая неполных последовательностей. Теперь выражение $b_i(\boldsymbol{o}_t)$, $i=\overline{1,N}$, $t=\overline{1,T}$ определено для всех $\boldsymbol{o}_t \in V^*$ в дискретном случае и для всех $\boldsymbol{o}_t \in R^*$ в непрерывном случае и формулы (7)-(12) расчёта прямых и обратных вероятностей можно расширить на случай неполных последовательностей.

Модифицированный алгоритм вычисления прямых вероятностей, используемый как при обучении СММ, так и при распознавании неполных последовательностей [72, 83]:

1) инициализация:

$$\alpha_1(i) = \begin{cases} \pi_i, & \boldsymbol{o}_1 = \varnothing \\ \pi_i b_i(\boldsymbol{o}_1), & \textit{иначе} \end{cases}, \qquad i = \overline{1, N};$$

2) индукция:

$$\alpha_{t+1}(i) = \begin{cases} \sum_{j=1}^{N} \alpha_t(j) a_{ji}, & \boldsymbol{o}_{t+1} = \varnothing \\ b_i(\boldsymbol{o}_{t+1}) \left[\sum_{j=1}^{N} \alpha_t(j) a_{ji} \right], & \textit{иначе} \end{cases}, \quad i = \overline{1, N},\ t = \overline{1, T-1};$$

3) завершение:

$$p(O \mid \lambda) = \sum_{i=1}^{N} \alpha_T(i).$$

Модифицированный алгоритм вычисления обратных вероятностей, используемый при обучении:

1) инициализация:

$$\beta_T(i) = 1, \quad i = \overline{1, N};$$

2) индукция:

$$\beta_t(i) = \begin{cases} \sum_{j=1}^{N} \beta_{t+1}(j) a_{ij}, & \boldsymbol{o}_{t+1} = \varnothing \\ \sum_{j=1}^{N} \beta_{t+1}(j) b_j(\boldsymbol{o}_{t+1}) a_{ij}, & \textit{иначе} \end{cases}, \quad i = \overline{1, N},\ t = \overline{1, T-1};$$

3) завершение (приведено для общности):

$$p(O \mid \lambda) = \begin{cases} \sum_{i=1}^{N} \beta_1(i)\, \pi_i, & \boldsymbol{o}_1 = \varnothing \\ \sum_{i=1}^{N} \beta_1(i)\, \pi_i\, b_i(\boldsymbol{o}_1), & \textit{иначе} \end{cases}.$$

Легко увидеть, что с помощью процедуры маргинализации можно решать задачу распознавания неполных последовательностей, например, с помощью метода максимума функции правдоподобия, описанного в п. 1.4, поскольку соответствующие формулы были скорррректированы для случая пропущенных наблюдений. Восстановления пропусков алгоритм маргинализации не предполагает.

3.2. Распознавание условно восстановленных неполных последовательностей

Предлагаемый метод включает в себя на первом этапе замещение пропусков в последовательности наиболее подходящими значениями с помощью процедуры, описанной в п. 2.3 главы 2, а затем распознавание её с помощью критерия максимума правдоподобия. Однако этот подход требует уточнения, поскольку для применения процедуры восстановления из пункта 2.3 требуется знание модели. Так как для последовательности неизвестна истинная метка класса, то имеет смысл условно восстанавливать неполную последовательность O с помощью той же СММ λ, по которой будет затем рассчитываться значение $P(O|\lambda)$.

В свою очередь, обучение СММ по последовательностям с пропусками можно осуществить, используя стандартные методы (например, алгоритм Баума-Велша), если предварительно восстановить данные последовательности. Для этого согласно п. 2.3 (глава 2) требуется знание модели. Если априорные знания отсутствуют, то модель нужно получить через процедуру обучения, например, используя подход обучения с маргинализацией, который описан в п. 4.1 (глава 4), а уже после восстановления можно попытаться уточнить модель проводя ее переобучение на восстановленных последовательностях. Эффективность подобного подхода необходимо проверить экспериментально. Очевидный недостаток такого подхода заключается в том, что обучение СММ необходимо проводить два раза.

Возможно также применять более сложную процедуру классификации, при которой одна и та же последовательность восстанавливается несколькими СММ, и затем вычисляется функция правдоподобия не только для той же СММ, которая восстановила последовательность, но и для всех остальных СММ, а финальная классификация делается по определенному набору правил. Продемонстрируем данную процедуру на примере двухклассовой классификации. Восстановим ее (все имеющиеся пропуски) с использованием модели λ_1. В результате получим полную последовательность $O_{(1)}$ ничем в статистическом смысле не отличающуюся от дру-

гих последовательностей класса 1. Дополнительно восстановим последовательность с помощью модели λ_2, получим последовательность $O_{(2)}$. Для последовательностей $O_{(1)}$ и $O_{(2)}$ вычислим правдоподобие по λ_1 и λ_2. Получим следующие величины $p(O_{(1)} \mid \lambda_1)$, $p(O_{(1)} \mid \lambda_2)$, $p(O_{(2)} \mid \lambda_2)$, $p(O_{(2)} \mid \lambda_1)$.

Возможны следующие ситуации:

1) $p(O_{(1)} \mid \lambda_1) > p(O_{(1)} \mid \lambda_2)$ и $p(O_{(2)} \mid \lambda_2) < p(O_{(2)} \mid \lambda_1)$. Тогда решение в пользу класса 1.

2) $p(O_{(1)} \mid \lambda_1) < p(O_{(1)} \mid \lambda_2)$ и $p(O_{(2)} \mid \lambda_2) > p(O_{(2)} \mid \lambda_1)$. Тогда решение в пользу класса 2 и это неправильное решение.

3) $p(O_{(1)} \mid \lambda_1) > p(O_{(1)} \mid \lambda_2)$ и $p(O_{(2)} \mid \lambda_2) > p(O_{(2)} \mid \lambda_1)$. Получаем разнонаправленный результат: в этом случае нужно вводить дополнительное правило. Сравниваем на каком классе правдоподобие больше, т. е. сравниваем $p(O_{(1)} \mid \lambda_1)$ и $p(O_{(2)} \mid \lambda_2)$. Если больше первое, то тогда решение в пользу класса 1. Если наоборот, то решение в пользу класса 2 [73].

3.3 Распознавание неполных последовательностей путём предварительного удаления пропусков

Пусть дана неполная последовательность O. Удалим из неё все пропуски, оставив только наблюдения, значения которых известны, в исходном порядке. После этого можно применять стандартные методы распознавания последовательностей, например, метод, основанный на критерии максимума функции правдоподобия (п. 1.4, глава 1). Такой метод борьбы с пропусками интуитивно является наиболее очевидным и крайне прост в реализации, однако, предсказуемо приводит к потере ценной информации – о том, что между известными наблюдениями было одно или несколько наблюдений с пропущенными значениями. Особенно наличие данной информации критично в случае, когда реальных физический процесс, описы-

ваемый СММ, порождает наблюдения через равные промежутки времени. Условимся также называть данный метод «склеиванием», так как после удаления пропусков между собой соединяются целые участки последовательности.

3.4 Исследование алгоритма распознавания неполных последовательностей, основанного на модифицированном алгоритме forward-backward

3.4.1 Оценка эффективности алгоритма распознавания неполных последовательностей, описываемых скрытыми марковскими моделями с дискретным распределением наблюдений, основанного на модифицированном алгоритме forward-backward

В данном вычислительном эксперименте проведено сравнение эффективности алгоритмов распознавания неполных последовательностей, описываемых СММ с дискретным распределением наблюдений. В реальных ситуациях может возникнуть необходимость решения задачи распознавания последовательностей с пропусками. В этом случае в качестве метрики для сравнения качества обучающих алгоритмов можно использовать количество верно распознанных последовательностей. Затрудним условия распознавания, выбрав достаточно близкие по параметрам две модели СММ с дискретным распределением наблюдений. Для этого рассмотрим две модели λ_1 и λ_2, различающиеся только матрицами вероятностей переходов $A = \begin{bmatrix} 0.1+\Delta A & 0.7-\Delta A & 0.2 \\ 0.2 & 0.2+\Delta A & 0.6-\Delta A \\ 0.8-\Delta A & 0.1+\Delta A & 0.1 \end{bmatrix}$ у первой модели $\lambda_1 \; - \; \Delta A = 0$, а у второй модели $\lambda_2 \; - \; \Delta A = 0.3$. Все остальные параметры у исходных моделей совпадают: число скрытых состояний $N = 3$, размерность алфавита наблюдаемых символов $M = 3$. Вектор распределения начального состояния: $\Pi = [1,0,0]$, матрица эмиссии: $B = \begin{bmatrix} 0.1 & 0.1 & 0.8 \\ 0.1 & 0.8 & 0.1 \\ 0.8 & 0.1 & 0.1 \end{bmatrix}$.

Проведено распознавание двух наборов по $K_C = 100$ тестовых последовательностей длиной $T_C = 100$ сгенерированных каждой из двух исходных моделей, соответственно, с различным процентом пропусков. Оценим эффективность распознавания таких последовательностей, если в классификаторе использованы исходные модели λ_1 и λ_2, по которым и проведена генерация тестовых последовательностей.

Рисунок 15 содержит результаты описанного выше эксперимента. Приведены усредненные величины для 100 запусков. Тип линии обозначает использованный метод классификации последовательностей с пропусками: сплошная – модифицированный алгоритм forward-backward (п. 3.1), штриховая – удаление пропусков из последовательностей (п. 3.3) и затем стандартный алгоритм распознавания

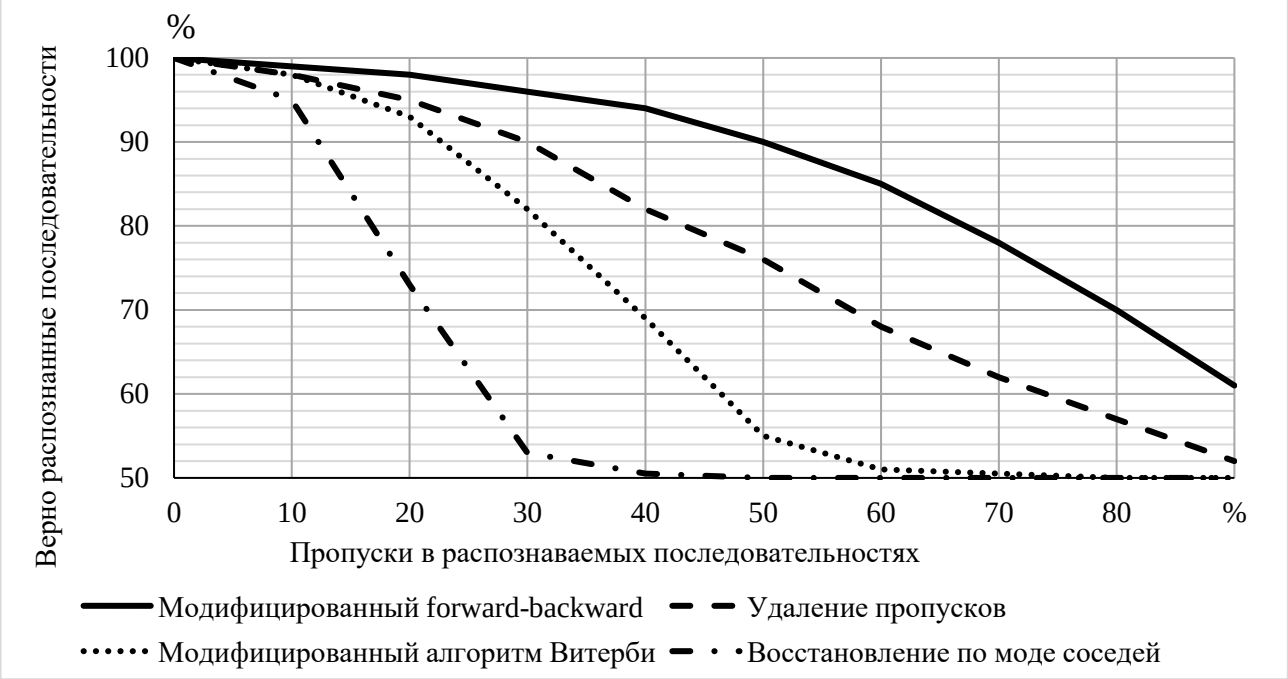

Рисунок 15 – Эффективность модифицированного алгоритма forward-backward для распознавания неполных последовательностей дискретных наблюдений

(п. 1.4), пунктирная – алгоритм восстановления последовательностей с пропусками с помощью модифицированного алгоритма Витерби и дальнейшее распознавание

стандартным алгоритмом (п. 3.2), штрихпунктирная – восстановление последовательностей с пропусками по моде соседних наблюдений (п. 2.4) и затем стандартный алгоритм распознавания (п. 1.4).

Как видно из диаграммы зависимостей на рисунке 15, эффективность распознавания с помощью метода маргинализации пропущенных наблюдений (модифицированный алгоритм forward-backward) наиболее высока. На втором месте алгоритм, основанный на склеивании последовательностей с пропусками и затем стандартное распознавание. Далее идёт алгоритм восстановления последовательностей с пропусками с помощью модифицированного алгоритма Витерби и затем стандартное распознавание. Худший результат показал метод восстановления последовательностей с пропусками по моде соседних наблюдений и затем стандартное распознавание.

Таким образом, метод классификации, основанный на модифицированном алгоритме forward-backward, позволяет достичь до 1.3 раза увеличения точности распознавания неполных последовательностей по сравнению с другими подходами [73].

3.4.2 Оценка эффективности алгоритма распознавания неполных последовательностей, описываемых скрытыми марковскими моделями с непрерывным распределением наблюдений, основанного на модифицированном алгоритме forward-backward

В данном вычислительном эксперименте проводилось сравнение различных подходов к распознаванию неполных последовательностей, описываемых СММ с непрерывным распределением наблюдений. В этом случае в качестве метрики для сравнения качества обучающих алгоритмов можно использовать количество верно распознанных последовательностей.

Затрудним условия распознавания, выбрав две, достаточно близкие по параметрам, модели СММ. В качестве истинных СММ были взяты модели λ_1 и λ_2 со

следующими характеристиками. Число скрытых состояний $N=3$, количество компонент в смесях $M=3$. Размерность векторов наблюдений $Z=2$. Модели λ_1 и λ_2, различались только матрицами вероятностей переходов $A=\begin{bmatrix} 0.1+\Delta A & 0.7-\Delta A & 0.2 \\ 0.2 & 0.2+\Delta A & 0.6-\Delta A \\ 0.8-\Delta A & 0.1+\Delta A & 0.1 \end{bmatrix}$. У первой модели λ_1 параметр $\Delta A=0$, а у второй модели λ_2 параметр $\Delta A=0.3$. Вектор распределения начального состояния: $\Pi=[1,0,0]$, веса компонент смесей $\left\{\tau_{im}, i=\overline{1,N}, m=\overline{1,M}\right\}=\begin{pmatrix} 0.3 & 0.4 & 0.3 \\ 0.3 & 0.4 & 0.3 \\ 0.3 & 0.4 & 0.3 \end{pmatrix}$ (номеру строки соответствует номер скрытого состояния, а номеру столбца – номер компоненты смеси), вектора математических ожиданий компонент смесей $\left\{\mu_{im}, i=\overline{1,N}, m=\overline{1,M}\right\}=\begin{pmatrix} (0 \quad 0)^T & (1 \quad 1)^T & (2 \quad 2)^T \\ (3 \quad 3)^T & (4 \quad 4)^T & (5 \quad 5)^T \\ (6 \quad 6)^T & (7 \quad 7)^T & (8 \quad 8)^T \end{pmatrix}$ (номеру строки соответствует номер скрытого состояния, а номеру столбца – номер компоненты смеси), все ковариационные матрицы компонент смесей $\left\{\Sigma_{im}, i=\overline{1,N}, m=\overline{1,M}\right\}$ были выбраны диагональными со значением 0.1 на диагонали. проводилось распознавание двух наборов, состоящих из $K_C=100$ тестовых последовательностей длиной $T_C=100$ без пропусков, сгенерированных каждой из двух исходных моделей соответственно. В качестве классификатора применялся критерий максимума функции правдоподобия (п. 1.4)

Рисунок 16 содержит результаты описанного выше эксперимента. Приведены средние значения после 100 запусков. Тип линии обозначает использованный метод классификации последовательностей с пропусками: сплошная – модифицированный алгоритм forward-backward (п. 3.1), штриховая – удаление пропусков из

последовательностей (п. 3.3) и затем применение стандартного алгоритма распознавания (п. 1.4), пунктирная – алгоритм восстановления последовательностей с пропусками с помощью модифицированного алгоритма Витерби и дальнейшее распознавание стандартным алгоритмом (п. 3.2), штрихпунктирная – восстановление последовательностей с пропусками по среднему арифметическому соседних наблюдений (п. 2.4) и затем применение стандартного алгоритм распознавания (п. 1.4).

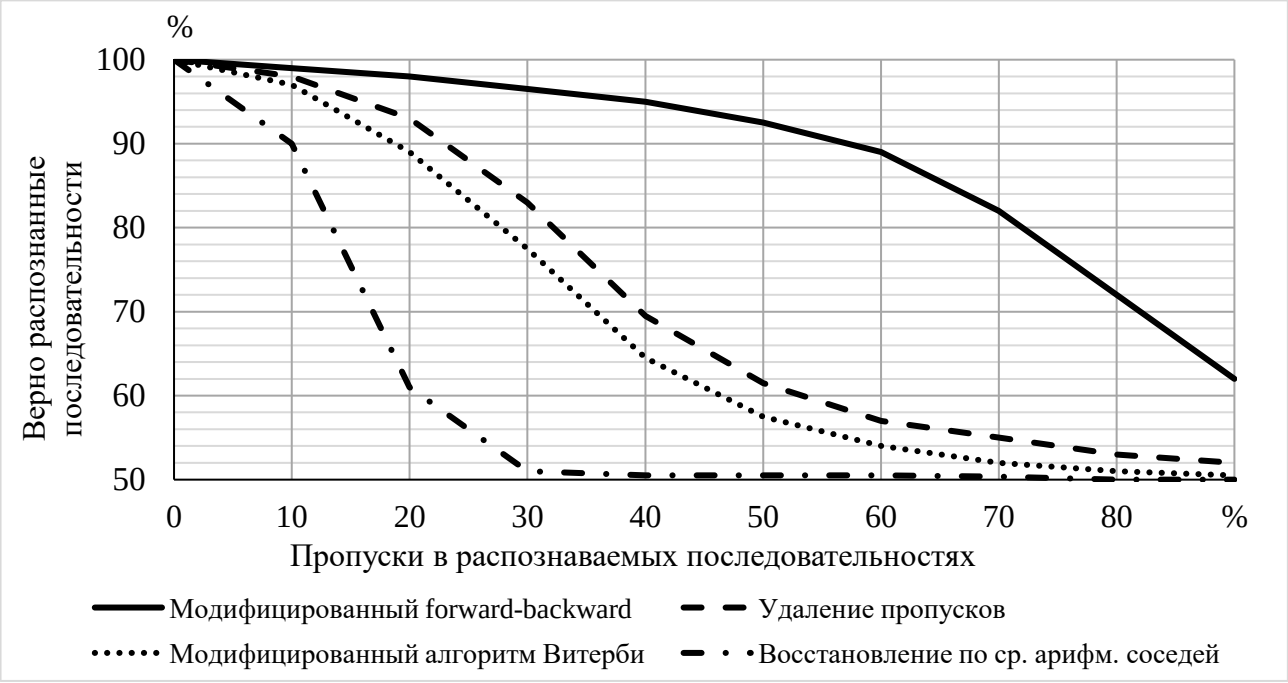

Рисунок 16 – Эффективность модифицированного алгоритма forward-backward для распознавания неполных последовательностей векторов вещественных чисел

Как видно из диаграммы зависимостей на рисунке 16, метод распознавания с помощью маргинализации пропущенных наблюдений (модифицированный алгоритм forward-backward) наиболее эффективен. На втором месте алгоритм, основанный на удалении пропусков из последовательностей с последующим стандартным распознаванием. Далее идёт алгоритм восстановления последовательностей с пропусками с помощью модифицированного алгоритма Витерби с последующим стандартным распознаванием. Худший результат показал алгоритм восстановления последовательностей с пропусками по среднему арифметическому соседних наблюдений с последующим стандартным распознаванием [75]. Вывод: новый метод на

основе модифицированного алгоритма forward-backward позволяет до 1.6 раз увеличить точность распознавания неполных последовательностей по сравнению с другими подходами.

3.5 Разработка методики идентификации личности по неполным данным двигательной активности при полной обучающей выборке

Данная практическая задача заключается в распознавании неполных последовательностей, генерируемых носимым устройством, анализирующим данные двигательной активности [84]. Задача идентификации пользователей по данным их двигательной активности имеет множество применений в современном мире. Например, смартфоны, носимые устройства и устройства интернета вещей могут быть оснащены функцией, которая определяет, что устройство используется кем-то отличным от их владельца и посылать сигнал предупреждения. Также, если носимое устройство используется разными членами семьи, то оно может автоматически определять, кто использует устройство в данный момент. Другие возможные применения включают в себя использование в военных целях (для определения того, что оборудование или снаряжение используется авторизованным пользователем) или криминалистике (установить личность человека, который использует устройство). К сожалению, данные с датчиков, как правило, подвержены искажениям и потере информации. Например, из-за временной неработоспособности датчика или воздействия внешнего шума. Именно поэтому алгоритмы анализа движения, применяемые в реальных условиях, должны быть устойчивы к таким неполадкам и должны уметь обрабатывать пропуски в данных оптимальным способом.

В качестве исходных данных использовалась та же выборка двигательной активности, что и в п. 2.7.1 (глава 2).

Для каждого участника эксперимента была обучена своя СММ. Каждая СММ имела N=3 скрытых состояния и M=3 компонент смесей из 3-хмерных (Z=3) нормальных распределений. Число скрытых состояний и компонент смесей были подобраны эмпирически таким образом, чтобы обеспечить наибольшую точность при

приемлемом времени расчёта. Каждая из больших последовательностей была разделена на подпоследовательности длиной T=100 (что соответствует примерно 3-м секундам наблюдения). Для обучения было использовано 75% случайно выбранных последовательностей из каждого класса. Тестирование обученных моделей проводилось на 25% оставшихся последовательностей. Для оценки эффективности классификации использовано количество верно распознанных последовательностей каждого класса.

Вышеупомянутая метрика была рассчитана для различных значений пропусков в тестовых последовательностях и разных алгоритмов борьбы с пропусками. Расположение пропусков было выбрано случайно в каждой из последовательностей. Рисунок 17 содержит результаты описанного выше эксперимента.

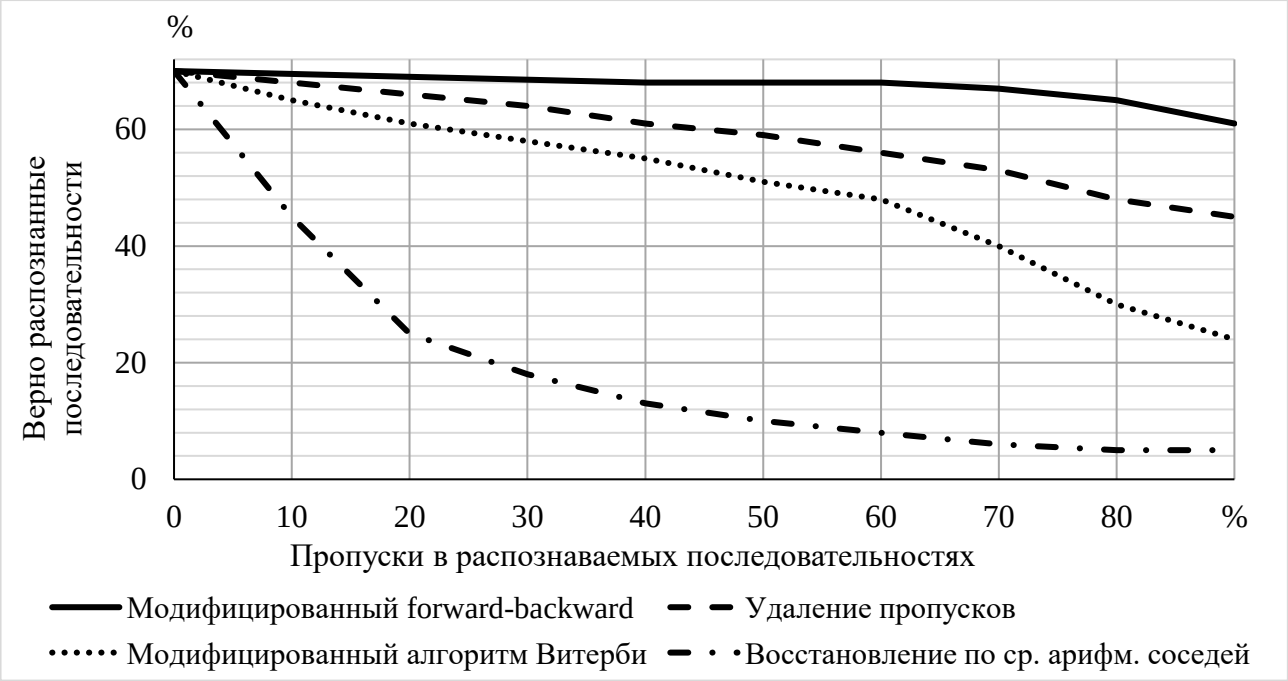

Рисунок 17 – Эффективность модифицированного алгоритма forward-backward для идентификации личности при неполных данных двигательной активности

Из диаграммы на рисунке 17 видно, что метод классификации, основанный на модифицированном алгоритме forward-backward, превосходит остальные подходы. Метод удаления пропусков в последовательностях и модифицированный алгоритм Витерби, после которых применялась стандартная процедура классификации, занимают 2 и 3 место соответственно. Алгоритм замены пропусков средним

арифметическим их соседей значительно уступает другим методам, а потому можно заключить, что он не пригоден для решения данной задачи.

В итоге, отталкиваясь от полученных результатов, можно рекомендовать использование алгоритма маргинализации для борьбы с пропусками при идентификации пользователей по данным их двигательной активности с помощью скрытых марковских моделей. Он превосходит другие методы, легко реализуем и не привносит дополнительных вычислительных затрат в процедуру классификации.

Выводы по третьей главе

В данной главе автором был предложен и научно обоснован метод распознавания последовательностей с пропусками, основанный на модифицированном алгоритме forward-backward. Преимущество предложенного метода по сравнению с другими подходами было подтверждено экспериментально. Он оказался эффективнее, чем следующие подходы: предварительное условное восстановлением последовательностей с помощью модифицированного алгоритма Витерби; удаление пропущенных наблюдений из неполных последовательностей; предварительное восстановление пропусков по моде в случае дискретного распределения наблюдений или по среднему арифметическому соседей в случае непрерывного распределения наблюдений.

Разработанный метод также был успешно применен для решения задачи идентификации личности по неполным данным двигательной активности. Таким образом, была доказана применимость разработанного метода к решению прикладной задачи, а также подтверждены результаты синтетических экспериментов.

ГЛАВА 4 РАЗРАБОТКА МЕТОДА ОБУЧЕНИЯ СКРЫТЫХ МАРКОВСКИХ МОДЕЛЕЙ ПО НЕПОЛНЫМ ПОСЛЕДОВАТЕЛЬНОСТЯМ

В данной главе описан новый метод, основанный на модифицированном алгоритме Баума-Велша, научно обосновано его применение для обучения скрытых марковских моделей по неполным последовательностям, а также проведен сравнительный анализ эффективности разработанного метода по отношению к другим методам анализа скрытых марковских моделей по неполным последовательностям [85].

4.1 Обучение скрытой марковской модели по неполным последовательностям с помощью модифицированного алгоритма Баума-Велша

Задача обучения в данном случае имеет ту же постановку, что и в п. 1.5.1 главы 1, за исключением того, что некоторые из обучающих последовательностей могут быть неполными. Для того, чтобы стало возможным применять алгоритм Баума-Велша (п. 1.5.1, формулы (14)-(22)) для обучения СММ по неполным последовательностям, некоторые формулы вычислений, входящие в алгоритм, необходимо модифицировать. Модификация алгоритма вычисления прямых-обратных вероятностей уже была приведена в п. 3.1 главы 3, поэтому будем считать величины $\alpha_t(i),\ i=\overline{1,N},\ t=\overline{1,T}$, $\beta_t(i),\ i=\overline{1,N},\ t=\overline{1,T}$ известными. Воспользуемся формулой эмиссии в случае пропущенного наблюдения, которая была получена в п. 2.1 для того, чтобы дополнить остальные формулы [86].

Модифицируем формулу (15):

$$\xi_t(i,j)=p(q_t=s_i,q_{t+1}=s_j\mid O,\hat{\lambda})=$$

$$=\begin{cases}\dfrac{\alpha_t(i)a_{ij}\beta_{t+1}(j)}{p(O\mid\hat{\lambda})}, & \boldsymbol{o}_{t+1}=\varnothing \\ \dfrac{\alpha_t(i)a_{ij}b_j(\boldsymbol{o}_{t+1})\beta_{t+1}(j)}{p(O\mid\hat{\lambda})}, & \textit{иначе}\end{cases},\qquad \begin{matrix} i,j=\overline{1,N}, \\ t=\overline{1,T-1}.\end{matrix} \tag{65}$$

Для случая обучения СММ с непрерывной плотностью распределения наблюдений необходимо внести изменения в формулу (16):

$$\gamma_t(i,m) = \begin{cases} \gamma_t(i)\,\tau_{im}, & \boldsymbol{o}_t = \varnothing \\ \gamma_t(i)\left[\dfrac{\tau_{im} g(\boldsymbol{o}_t, \mu_{im}, \Sigma_{im})}{b_i(\boldsymbol{o}_t)}\right], & \textit{иначе.} \end{cases} \quad (66)$$

Кроме того, формулы оценивания матриц ковариаций нормальных распределений, входящих в смеси, изменятся следующим образом:

$$\hat{\Sigma}'_{im} = \frac{\sum_{k=1}^{K}\sum_{\substack{t=1 \\ \boldsymbol{o}_t \neq \varnothing}}^{T^k-1} \gamma_t^{(k)}(i,m)(\boldsymbol{o}_t^{(k)} - \hat{\mu}'_{im})(\boldsymbol{o}_t^{(k)} - \hat{\mu}'_{im})^T}{\sum_{k=1}^{K}\sum_{\substack{t=1 \\ \boldsymbol{o}_t \neq \varnothing}}^{T^k-1} \gamma_t^{(k)}(i,m)}. \quad (67)$$

Как можно заметить, отличие состоит в том, что в данной формуле суммируются только те компоненты, которым соответствуют наблюдения, не являющиеся пропусками.

Таким образом, с помощью модифицированного алгоритма Баума-Велша становится возможным проводить обучение СММ по неполным последовательностям [87].

4.2 Обучение скрытой марковской модели по неполным последовательностям, восстановленным с помощью модифицированного алгоритма Витерби

В свою очередь, обучение СММ по последовательностям с пропусками можно осуществить, используя стандартные методы (например, алгоритм Баума-Велша), если предварительно восстановить данные последовательности. Для этого согласно п. 2.3 (глава 2) требуется знание модели. Если априорные знания отсутствуют, то модель нужно получить через процедуру обучения, например, с помощью модифицированного алгоритма Баума-Велша из п. 4.1, а уже после восстановления можно попытаться уточнить модель проводя ее переобучение на вос-

становленных последовательностях. Эффективность подобного подхода необходимо проверить экспериментально. Очевидный недостаток такого подхода заключается в том, что обучение СММ необходимо проводить два раза [88].

4.3 Обучение скрытой марковской модели по неполным последовательностям путём удаления пропусков

Пусть дано множество неполных обучающих последовательностей $O^* = \{O^1, O^2, ..., O^K\}$, где K – это количество последовательностей. Удалим из каждой из них все пропуски, оставив только наблюдения, значения которых известны, в исходном порядке. После этого можно применять стандартный алгоритм обучения СММ, например, алгоритм Баума-Велша, описанный в п. 1.5.1. Ранее мы условились называть данный метод «склеиванием» [89].

4.4 Исследование модифицированного алгоритма Баума-Велша обучения скрытой марковской модели

4.4.1 Оценка эффективности модифицированного алгоритма Баума-Велша обучения скрытой марковской модели с дискретным распределением наблюдений

В данном вычислительном эксперименте проведено сравнение различных подходов к обучению СММ по последовательностям, содержащим пропуски. В качестве истинной СММ была взята модель λ со следующими характеристиками. Число скрытых состояний $N = 3$, размерность алфавита наблюдаемых символов $M = 3$. Вектор распределения начального состояния: $\Pi = [1,0,0]$, матрица вероятностей переходов: $A = \begin{bmatrix} 0.1 & 0.7 & 0.2 \\ 0.2 & 0.2 & 0.6 \\ 0.8 & 0.1 & 0.1 \end{bmatrix}$, матрица эмиссии: $B = \begin{bmatrix} 0.1 & 0.1 & 0.8 \\ 0.1 & 0.8 & 0.1 \\ 0.8 & 0.1 & 0.1 \end{bmatrix}$. С помощью процесса, описываемого данной СММ, было сгенерировано $K = 100$ обучаю-

щих последовательностей $\{O^1, O^2, ..., O^K\}$ длиной $T = 100$. В ходе исследования изменялось количество пропусков в обучающих последовательностях $\{O^1, O^2, ..., O^K\}$, которые использовались для нахождения оценки параметров модели – λ. Пропуски генерировались случайным образом, причем в различных местах каждой последовательности. Выход из итерационного процесса обучения осуществлялся по сходимости.

При изменении количества пропусков фиксировалось изменение следующих величин. Во-первых, фиксировалось значение логарифма функции правдоподобия того, что обученная модель сгенерировала исходные обучающие последовательности (без пропусков), т. е. $\ln p\left(\{O^1, O^2, ..., O^K\} \mid \lambda\right)$. Во-вторых, фиксировалось расстояние, основанное на симметричной разности логарифмов правдоподобия, между истинной и обученной моделью. Это расстояние вычисляется по следующей формуле:

$$D_s = \frac{D(\lambda, \lambda) + D(\lambda, \lambda)}{2}, \qquad (68)$$

где $D(\lambda_1, \lambda_2) = \frac{1}{T}\left| \ln p\left(O^2 \middle| \lambda_1\right) - \ln p\left(O^2 \middle| \lambda_2\right) \right|$, а O^2 – последовательность, порождённая λ_2. Данная метрика позволяет более адекватным образом сравнить две СММ, нежели норма разности параметров [24]. Для расчётов по формуле (68) генерировалось $K_D = 100$ последовательностей длиной $T_D = 500$ для каждой СММ и брался средний результат.

Рисунок 18 содержит результаты описанного выше эксперимента при использовании в качестве метрики качества обучения логарифм функции правдоподобия. Приведены усредненные значения для 100 проведённых экспериментов. Рисунок 19 содержит результаты эксперимента при использовании в качестве метрики расстояния, основанного на правдоподобии между истинной моделью и её оценкой. Здесь также даны усредненные значения для 100 проведённых экспериментов.

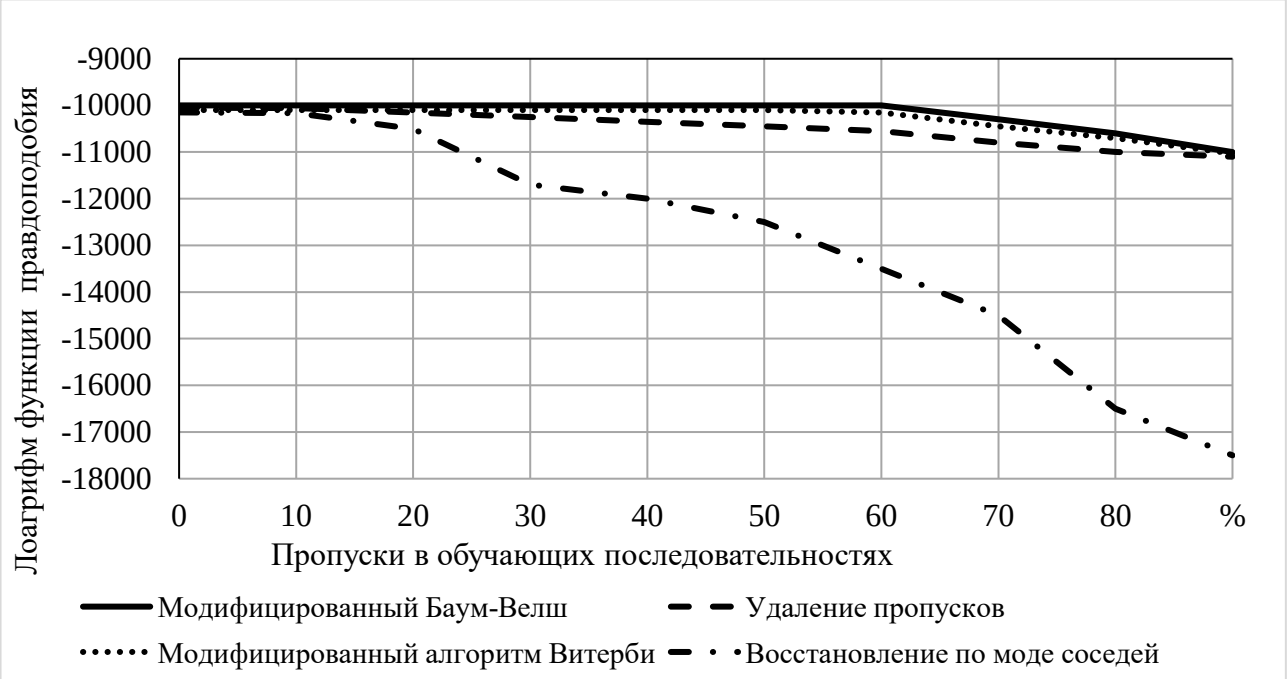

Рисунок 18 – Эффективность модифицированного алгоритма Баума-Велша при метрике качества обучения: логарифм функции правдоподобия (дискретное распределение наблюдений)

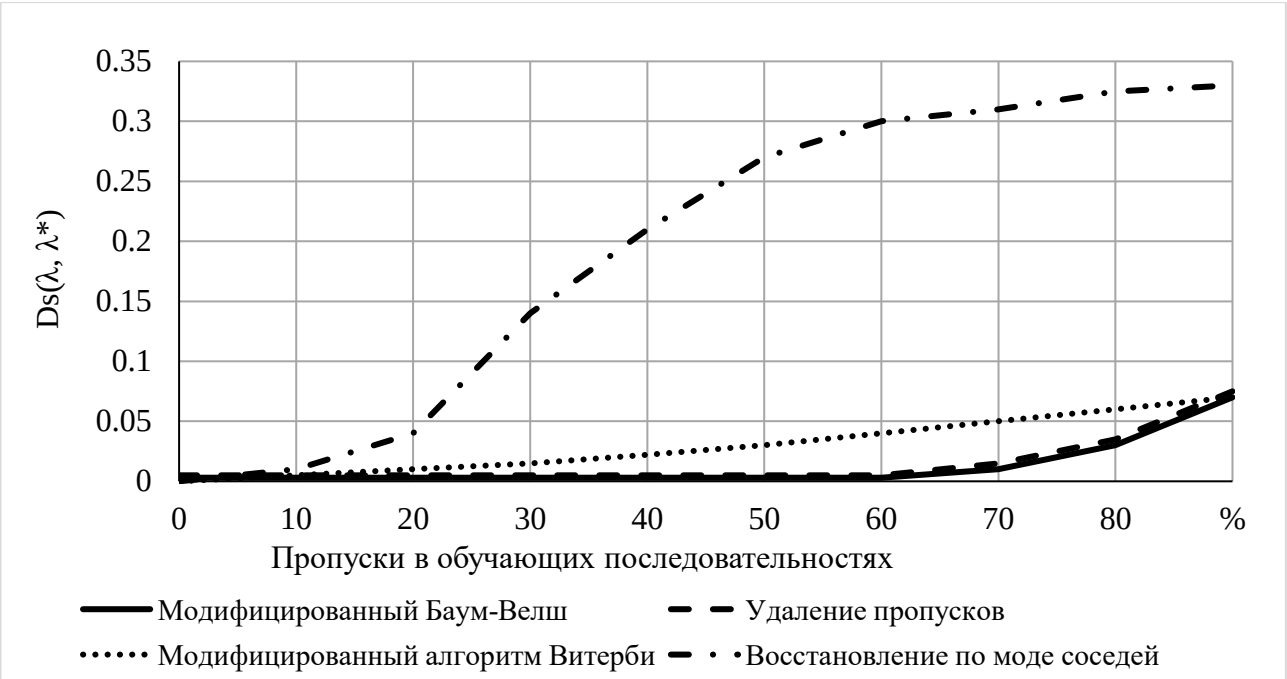

Рисунок 19 – Эффективность модифицированного алгоритма Баума-Велша при метрике качества обучения: расстояние, основанное на правдоподобии между истинной моделью и её оценкой (дискретное распределение наблюдений)

На обоих рисунках представлены графики зависимости величин от пропусков в обучающих последовательностях. Тип линии обозначает использованный метод обучения: сплошная – модифицированный алгоритм Баума-Велша (п. 4.1), штриховая – удаление пропусков из последовательностей (п. 4.3) и затем использование стандартного алгоритма Баума-Велша (п. 1.5.1), пунктирная – восстановление последовательностей с пропусками с помощью модифицированного алгоритма Витерби (п. 2.3), и затем использование стандартного алгоритма Баума-Велша (п. 1.5.1), штрихпунктирная – восстановление последовательностей с пропусками по моде соседних наблюдений (п. 2.4) и затем использование стандартного алгоритма Баума-Велша (п. 1.5.1).

Как видно из вышеприведенных графиков на рисунках 18 и 19, алгоритм, использующий модифицированный алгоритм Баума-Велша и алгоритм, производящий предварительное восстановление пропусков по модифицированному алгоритму Витерби, очень близки по эффективности. Несколько меньшую эффективность демонстрирует алгоритм обучения, основанный на склеивании последовательностей с пропусками. Метод, основанный на восстановлении пропусков по моде ближайших соседей, показывает неудовлетворительные результаты.

Важнейшим показателем работоспособности алгоритмов обучения является использования их для построения классификаторов на основе полученных моделей. В этом случае в качестве метрики для сравнения эффективности обучающих алгоритмов можно использовать количество правильно распознанных последовательностей. Затрудним условия распознавания, выбрав достаточно близкие по параметрам две модели СММ. Для этого рассмотрим две модели λ_1 и λ_2, различающиеся только матрицами вероятностей переходов $A = \begin{bmatrix} 0.1+\Delta A & 0.7-\Delta A & 0.2 \\ 0.2 & 0.2+\Delta A & 0.6-\Delta A \\ 0.8-\Delta A & 0.1+\Delta A & 0.1 \end{bmatrix}$ у первой модели $\lambda_1 \; - \; \Delta A = 0$ (т.е. она совпадает с матрицей из предыдущего эксперимента), а у второй модели $\lambda_2 \; - \; \Delta A = 0.3$. Все остальные параметры у исходных моделей совпадают и равны параметрам модели,

использованной в предыдущем эксперименте. Оценка каждой из двух моделей проводилось по набору из $K=100$ обучающих последовательностей длиной $T=100$, сгенерированному соответствующей истинной моделью. После оценки проводилось распознавание двух наборов по $K_C=100$ тестовых последовательностей длиной $T_C=100$ без пропусков, сгенерированных каждой из двух исходных моделей соответственно. В качестве классификатора применялся алгоритм максимума логарифма правдоподобия (п. 1.4). Рисунок 20 содержит результаты данного эксперимента. На графике приведены усредненные значения после 100 запусков.

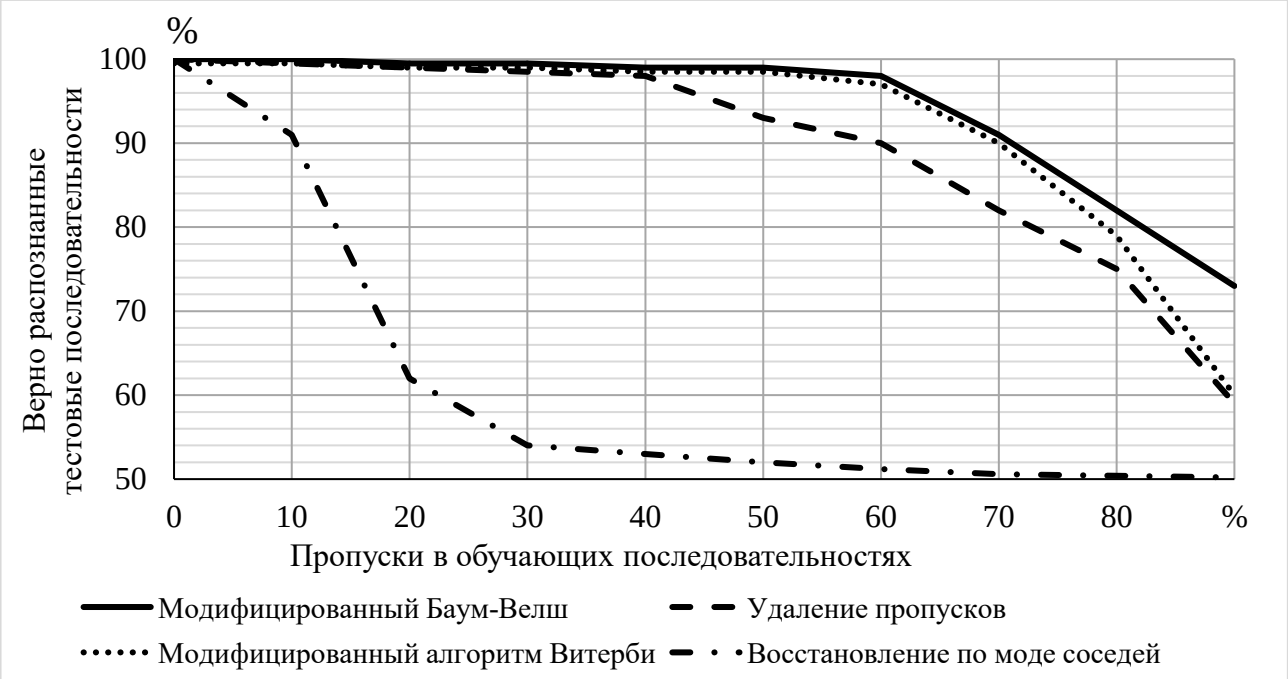

Рисунок 20 – Эффективность модифицированного алгоритма Баума-Велша для построения классификаторов целых последовательностей (дискретное распределение наблюдений)

Как видно из графика зависимостей на рисунке 20, обучение с помощью модифицированного алгоритма Баума-Велша обеспечивает наилучшие дискриминационные свойства полученных моделей. Модели, обученные алгоритмом с использованием восстановления пропусков по модифицированному алгоритму Витерби, показывают чуть меньшее количество правильно распознанных последовательностей. Хуже работает алгоритм обучения, основанный на удалении пропусков из последовательностей. Метод, основанный

на восстановлении пропусков по моде ближайших соседей, в очередной раз показывает неудовлетворительные результаты, даже несмотря на то, что он был оптимизирован по числу соседей.

В реальных ситуациях может возникнуть необходимость решения задачи распознавания не только целых последовательностей, но и последовательностей с пропусками. В предыдущем эксперименте было рассмотрено, как меняется эффективность распознавания неполных последовательностей, если в классификаторе использовать исходные модели, по которым проводилась генерация этих последовательностей. В этот раз рассмотрим наиболее реалистичный случай, когда СММ, обученные на последовательностях с пропусками, будут применяться для классификации подобных «дефектных» последовательностей. Данное исследование было проведено таким же образом, как и описанный выше эксперимент по распознаванию последовательностей без пропусков с помощью моделей, обученных на последовательностях с пропусками. Отличие состояло в том, что в распознаваемых последовательностях теперь появлялись пропуски, причём количество пропусков в распознаваемых и в обучающих последовательностях было равным. Фиксировалось количество правильно распознанных последовательностей при изменении пропусков в обучающих и распознаваемых последовательностях.

Рисунок 21 содержит результаты данного эксперимента. На графике приведены усредненные значения для 100 экспериментов. Тип линии обозначает использованный метод обучения и распознавания: сплошная – обучение с помощью модифицированного алгоритма Баума-Велша (п. 4.1) и распознавание с помощью модифицированного алгоритма forward-backward и критерия МФП (п. 3.1), штриховая – обучение и распознавание стандартными алгоритмами путём предварительного исключения пропусков из последовательностей (п. 4.3), пунктирная – обучение и распознавание стандартными алгоритмами путём предварительного восстановления неполных последовательностей с помощью модифицированного алгоритма

Витерби (п. 2.3), штрихпунктирная – обучение и распознавание стандартными алгоритмами путём предварительного восстановления последовательностей с пропусками по моде соседних наблюдений (п. 2.4).

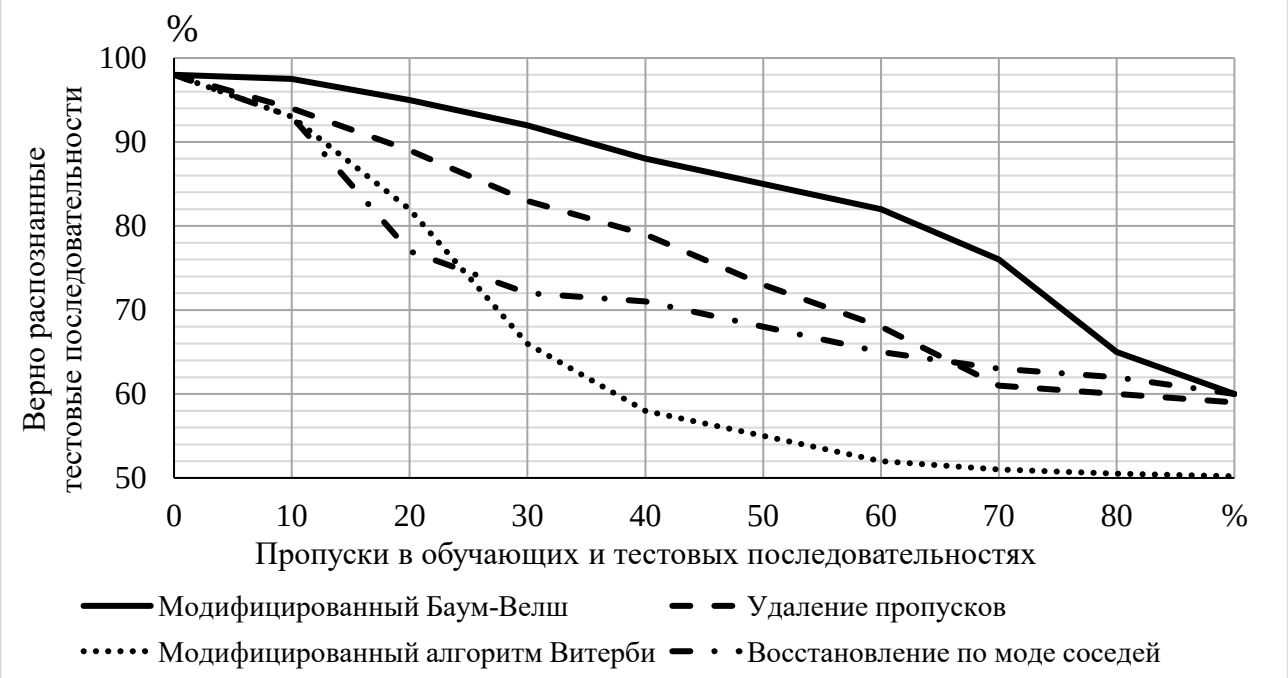

Рисунок 21 – Эффективность модифицированного алгоритма Баума-Велша для построения классификаторов неполных последовательностей (дискретное распределение наблюдений)

Как видно из рисунка 21, наилучший результат демонстрирует метод распознавания последовательностей на основе модифицированного алгоритма forward-backward, который использовал СММ, обученные с помощью модифицированного алгоритма Баума-Велша. Его точность распознавания неполных последовательностей до 1.2 раз выше, чем точность алгоритмов, предполагающих обучение и распознавание с помощью восстановления или удаление пропусков [90].

4.4.2 Оценка эффективности модифицированного алгоритма Баума-Велша обучения скрытой марковской модели с непрерывным распределением наблюдений

В данном вычислительном эксперименте проводилось сравнение эффективности нескольких алгоритмов обучения СММ с непрерывным распределением

наблюдений по неполным последовательностям. В качестве истинной СММ была взята модель λ со следующими характеристиками. Число скрытых состояний $N=3$, количество компонент в смесях $M=3$. Размерность векторов наблюдений $Z=2$. Вектор распределения начального состояния: $\Pi=[1,0,0]$, матрица вероятностей переходов: $A=\begin{bmatrix}0.1 & 0.7 & 0.2\\ 0.2 & 0.2 & 0.6\\ 0.8 & 0.1 & 0.1\end{bmatrix}$, веса компонент смесей $\left\{\tau_{im}, i=\overline{1,N}, m=\overline{1,M}\right\}=\begin{pmatrix}0.3 & 0.4 & 0.3\\ 0.3 & 0.4 & 0.3\\ 0.3 & 0.4 & 0.3\end{pmatrix}$ (номеру строки соответствует номер скрытого состояния, а номеру столбца – номер компоненты смеси), вектора математических ожиданий компонент смесей $\left\{\mu_{im}, i=\overline{1,N}, m=\overline{1,M}\right\}=\begin{pmatrix}(0 \quad 0)^T & (1 \quad 1)^T & (2 \quad 2)^T\\ (3 \quad 3)^T & (4 \quad 4)^T & (5 \quad 5)^T\\ (6 \quad 6)^T & (7 \quad 7)^T & (8 \quad 8)^T\end{pmatrix}$ (номеру строки соответствует номер скрытого состояния, а номеру столбца – номер компоненты смеси), все ковариационные матрицы компонент смесей $\left\{\Sigma_{im}, i=\overline{1,N}, m=\overline{1,M}\right\}$ были выбраны единичными. По заданной СММ было сгенерировано $K=100$ обучающих последовательностей $\left\{O^1, O^2, ..., O^K\right\}$ длиной $T=100$.

В ходе исследования менялось количество пропусков в обучающих последовательностях. Пропуски распределялись случайным образом в каждой последовательности. Выход из итерационного процесса обучения осуществлялся по сходимости. При изменении количества пропусков фиксировалось изменение значения логарифма функции правдоподобия того, что обученная модель сгенерировала целые обучающие последовательности, т. е. $\ln p\left(\left\{O^1, O^2, ..., O^K\right\} \mid \lambda\right)$, а также расстояние, основанное на симметричной разности логарифмов правдоподобия,

между истинной и обученной моделью (формула (68)). Для расчётов по формуле было сгенерировано $K_D = 100$ последовательностей длиной $T_D = 500$.

Рисунок 22 содержит результаты описанного выше эксперимента при использовании в качестве метрики качества обучения логарифм функции правдоподобия.

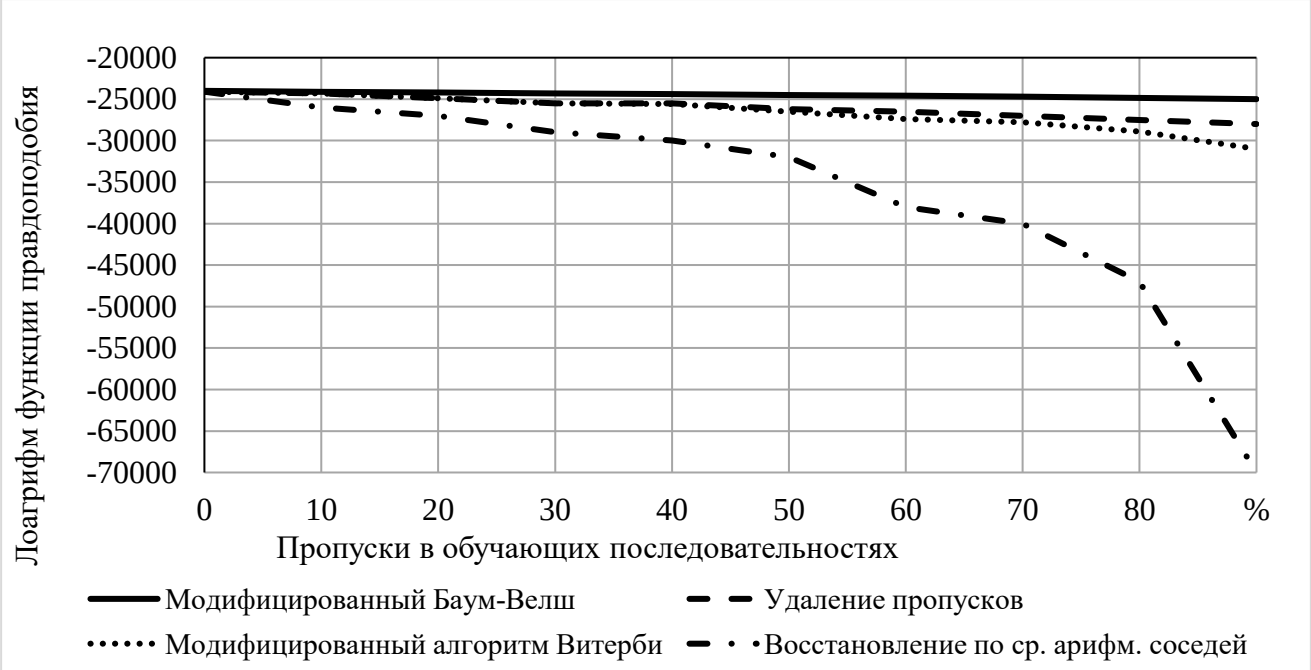

Рисунок 22 – Эффективность модифицированного алгоритма Баума-Велша при метрике качества обучения: логарифм функции правдоподобия (дискретное распределение наблюдений)

На

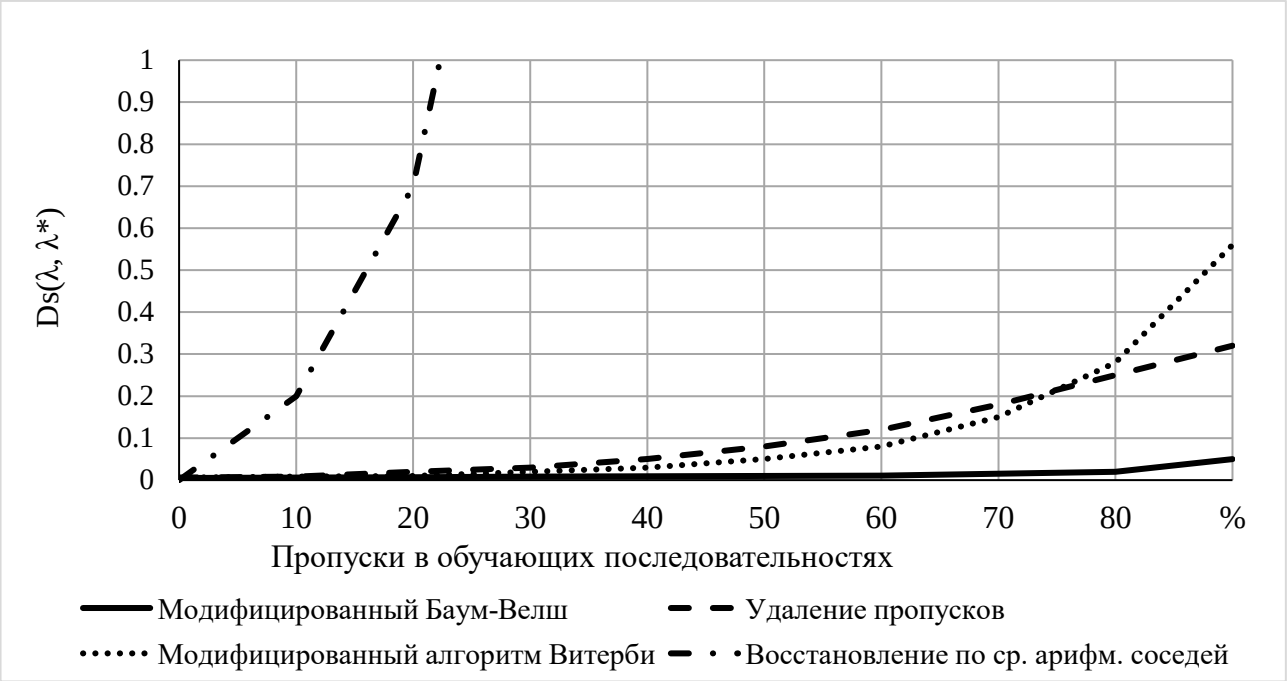

Рисунок 23 результаты эксперимента при метрике: расстояние, основанное на правдоподобии между истинной моделью и её оценкой. Приведены средние значения после 100 проведённых экспериментов. Тип линии обозначает использованный метод обучения: сплошная – модифицированный алгоритм Баума-Велша (п. 4.1), штриховая – удаление пропусков из последовательностей (п. 4.3) и затем использование стандартного алгоритма Баума-Велша (п. 1.5.1), пунктирная – восстановление последовательностей с пропусками с помощью модифицированного алгоритма Витерби (п. 2.1) и затем использование стандартного алгоритма Баума-Велша (п. 1.5.1), штрихпунктирная – восстановление последовательностей с пропусками по среднему арифметическому соседних наблюдений (п. 2.5) и затем использование стандартного алгоритма Баума-Велша (п. 1.5.1).

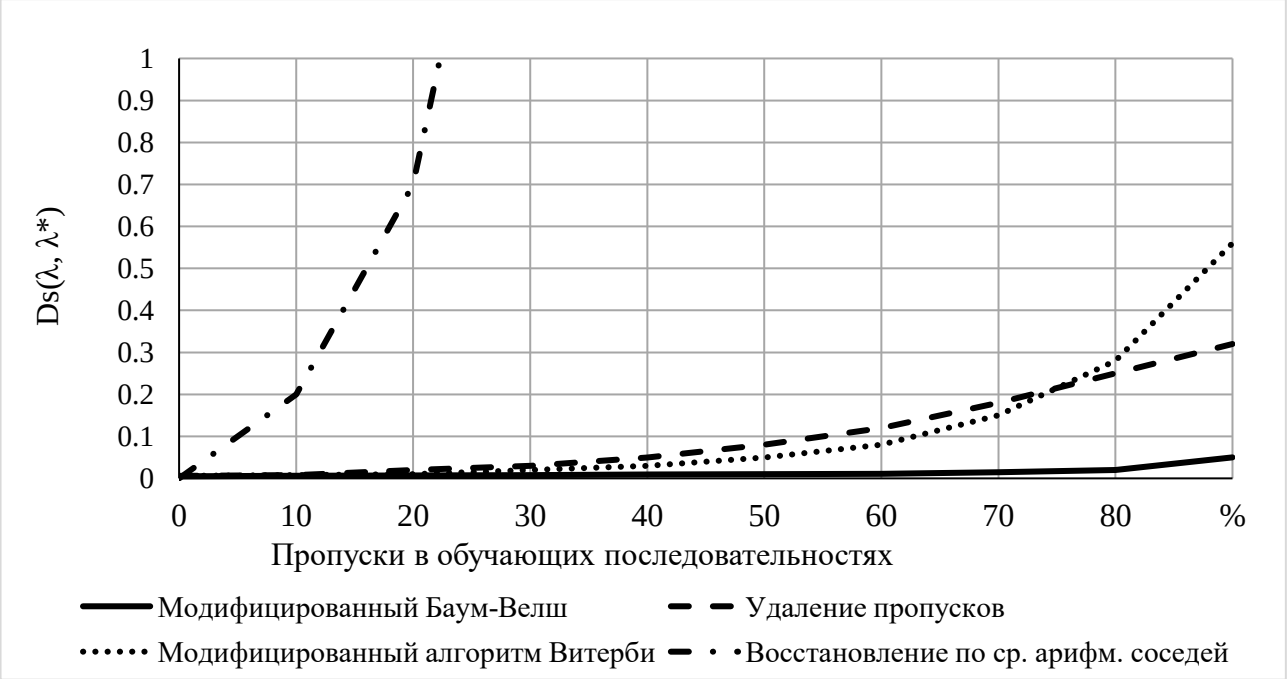

Рисунок 23 – Эффективность модифицированного алгоритма Баума-Велша при метрике качества обучения: расстояние, основанное на правдоподобии между истинной моделью и её оценкой (непрерывное распределение наблюдений)

Как видно из вышеприведенных графиков на рисунках 22 и 23, модифицированный алгоритм Баума-Велша с маргинализацией пропусков, показывает наилучшие результаты. Алгоритм, использующий восстановление пропусков по модифицированному алгоритму Витерби и алгоритм обучения, основанный на склеивании последовательностей с пропусками, очень близки по эффективности. Метод, основанный на восстановлении пропусков по среднему арифметическому ближайших соседей, показывает неудовлетворительные результаты.

Так же, как в п.4.4.1, для проверки качества алгоритмов обучения рассмотрим эффективность построенных с его помощью моделей при решении задачи классификации. В качестве метрики для сравнения качества обучающих алгоритмов использовано количество правильно распознанных последовательностей. Затрудним условия распознавания, выбрав достаточно близкие по параметрам две модели СММ. Для этого рассмотрим две модели λ_1 и λ_2, различающиеся только матрицами

вероятностей переходов $A = \begin{bmatrix} 0.1+\Delta A & 0.7-\Delta A & 0.2 \\ 0.2 & 0.2+\Delta A & 0.6-\Delta A \\ 0.8-\Delta A & 0.1+\Delta A & 0.1 \end{bmatrix}$ у первой модели λ_1 – $\Delta A = 0$ (т.е. она совпадает с матрицей из предыдущего эксперимента), а у второй модели λ_2 – $\Delta A = 0.3$. Остальные параметры у исходных моделей совпадают и равны параметрам модели, использованной в предыдущем эксперименте. Нахождение оценок каждой из двух моделей проводилось по набору из $K = 100$ обучающих последовательностей длиной $T = 100$, сгенерированному соответствующей истинной моделью. После нахождения оценок проводилось распознавание двух наборов по $K_C = 100$ тестовых последовательностей длиной $T_C = 100$ без пропусков, сгенерированных каждой из двух исходных моделей соответственно. В качестве классификатора применялся алгоритм максимума логарифма правдоподобия (п. 1.4). Рисунок 24 содержит результаты данного эксперимента. На графике приведены усредненные значения после 100 запусков.

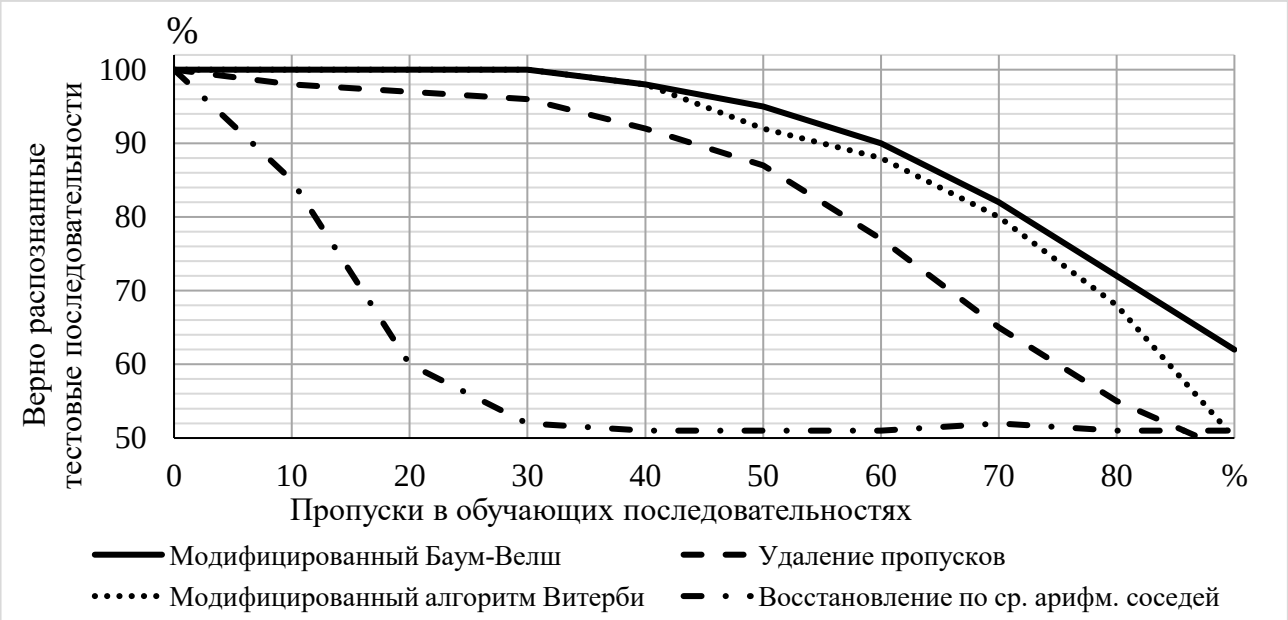

Рисунок 24 – Эффективность модифицированного алгоритма Баума-Велша для построения классификаторов целых последовательностей (непрерывное распределение наблюдений)

Как видно из графика на рисунке 24, обучение с помощью модифицированного алгоритма Баума-Велша (маргинализации пропущенных наблюдений) обеспечивает наилучшие дискриминационные свойства полученных моделей. Модели, обученные алгоритмом с использованием восстановления пропусков по модифицированному алгоритму Витерби, показывают чуть меньший процент правильно распознанных последовательностей. Чуть больше уступает алгоритм обучения, основанный на склеивании последовательностей с пропусками. Метод, основанный на восстановлении пропусков по моде ближайших соседей, в очередной раз показывает неудовлетворительные результаты, даже несмотря на то, что он был оптимизирован по числу соседей.

В реальных ситуациях может возникнуть необходимость решения задачи распознавания не только целых последовательностей, но и последовательностей с пропусками. В п. 3.5.1 мы рассмотрели, как меняется эффективность распознавания неполных последовательностей, если в классификаторе использовать исходные модели, по которым проводилась генерация этих последовательностей. В этот раз рассмотрим наиболее реалистичный на наш взгляд случай, когда СММ, обученные на последовательностях с пропусками, будут применяться для классификации подобных «дефектных» последовательностей. Данное исследование было проведено таким же образом, как и описанный выше эксперимент по распознаванию последовательностей без пропусков с помощью моделей, обученных на последовательностях с пропусками. Отличие состояло в том, что в распознаваемых последовательностях теперь появлялись пропуски, причём количество пропусков в распознаваемых и обучающих последовательностях были равны. Фиксировались правильно распознанные последовательности при изменении количества пропусков в обучающих и распознаваемых последовательностях.

Рисунок 25 содержит результаты данного эксперимента. На графике приведены средние значения после 100 проведений эксперимента. Тип линии обозначает использованный метод обучения и распознавания: сплошная – обучение с помощью модифицированного алгоритма Баума-Велша (п. 4.1) и распознавание с помощью модифицированного алгоритма forward-backward и критерия МФП (п. 3.1),

штриховая – обучение и распознавание стандартными алгоритмами путём предварительного исключения пропусков из последовательностей (п. 4.3), пунктирная – обучение и распознавание стандартными алгоритмами путём предварительного восстановления неполных последовательностей с помощью модифицированного алгоритма Витерби (п. 2.3), штрихпунктирная – обучение и распознавание стандартными алгоритмами путём предварительного восстановления последовательностей с пропусками по среднему арифметическому соседних наблюдений (п. 2.4).

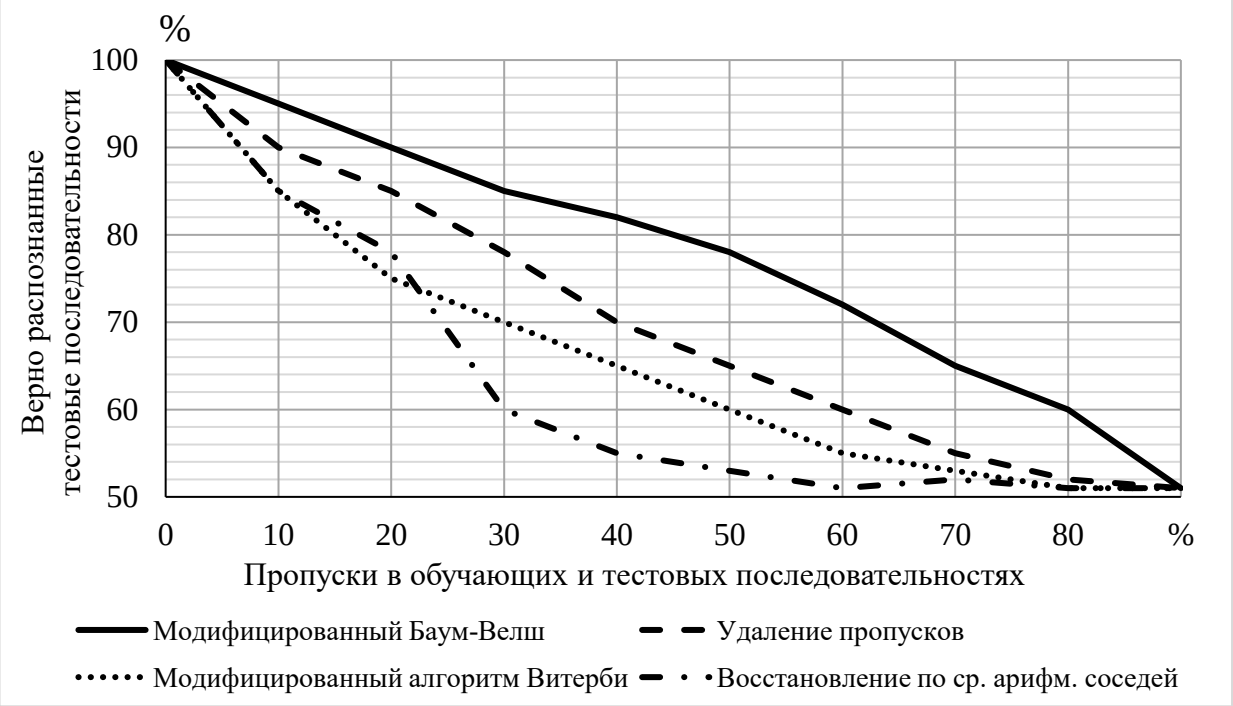

Рисунок 25 – Эффективность модифицированного алгоритма Баума-Велша для построения классификаторов неполных последовательностей (непрерывное распределение наблюдений)

Как видно из рисунка 25, наивысшую эффективность демонстрирует алгоритм распознавания последовательностей на основе модифицированного алгоритма forward-backward, который использовал СММ, обученные с помощью модифицированного алгоритма Баума-Велша. Это позволяет до 1.2 раз увеличить качество распознавания неполных последовательностей при наличии пропусков в обучающих последовательностях по сравнению с другими подходами к обучению и распознаванию [91].

4.5 Разработка методики идентификации личности по неполным данным двигательной активности при неполной обучающей выборке

Данная практическая задача заключается в обучении СММ на неполных последовательностях, генерируемых носимым устройством, анализирующим данные двигательной активности, и дальнейшего распознавания неполных последовательностей [84].

В качестве исходных данных использовалась та же выборка двигательной активности, что и в п. 2.7.1. Для каждого участника эксперимента была обучена своя СММ. Каждая СММ имела N=3 скрытых состояния и M=3 компонент смесей из 3-хмерных (Z=3) нормальных распределений. Число скрытых состояний и компонент смесей были подобраны эмпирически таким образом, чтобы обеспечить наибольшую точность при приемлемом времени расчёта. Каждая из больших последовательностей была разделена на подпоследовательности длиной T=100 (что соответствует примерно 3-м секундам наблюдения). Для обучения было использовано 75% случайно выбранных последовательностей из каждого класса. Тестирование обученных моделей проводилось на 25% оставшихся последовательностей. Для качества классификации была взята усредненная величина правильно распознанных последовательностей из каждого класса. Вышеупомянутая метрика была рассчитана для разного количества пропусков в обучающих и тестовых последовательностях и разных алгоритмов обучения СММ на неполных последовательностях. Расположение пропусков было выбрано случайно в каждой из обучающих последовательностей.

Рисунок 26 содержит результаты данного эксперимента. На графике приведены средние значения после 100 проведений эксперимента. Тип линии обозначает использованный метод обучения и распознавания: сплошная – обучение с помощью модифицированного алгоритма Баума-Велша (п. 4.1) и распознавание с помощью модифицированного алгоритма forward-backward и критерия МФП (п. 3.1), штриховая – обучение и распознавание стандартными алгоритмами путём предварительного исключения пропусков из последовательностей (п. 4.3), пунктирная –

обучение и распознавание стандартными алгоритмами путём предварительного восстановления неполных последовательностей с помощью модифицированного алгоритма Витерби (п. 2.3), штрихпунктирная – обучение и распознавание стандартными алгоритмами путём предварительного восстановления последовательностей с пропусками по среднему арифметическому соседних наблюдений (п. 2.4).

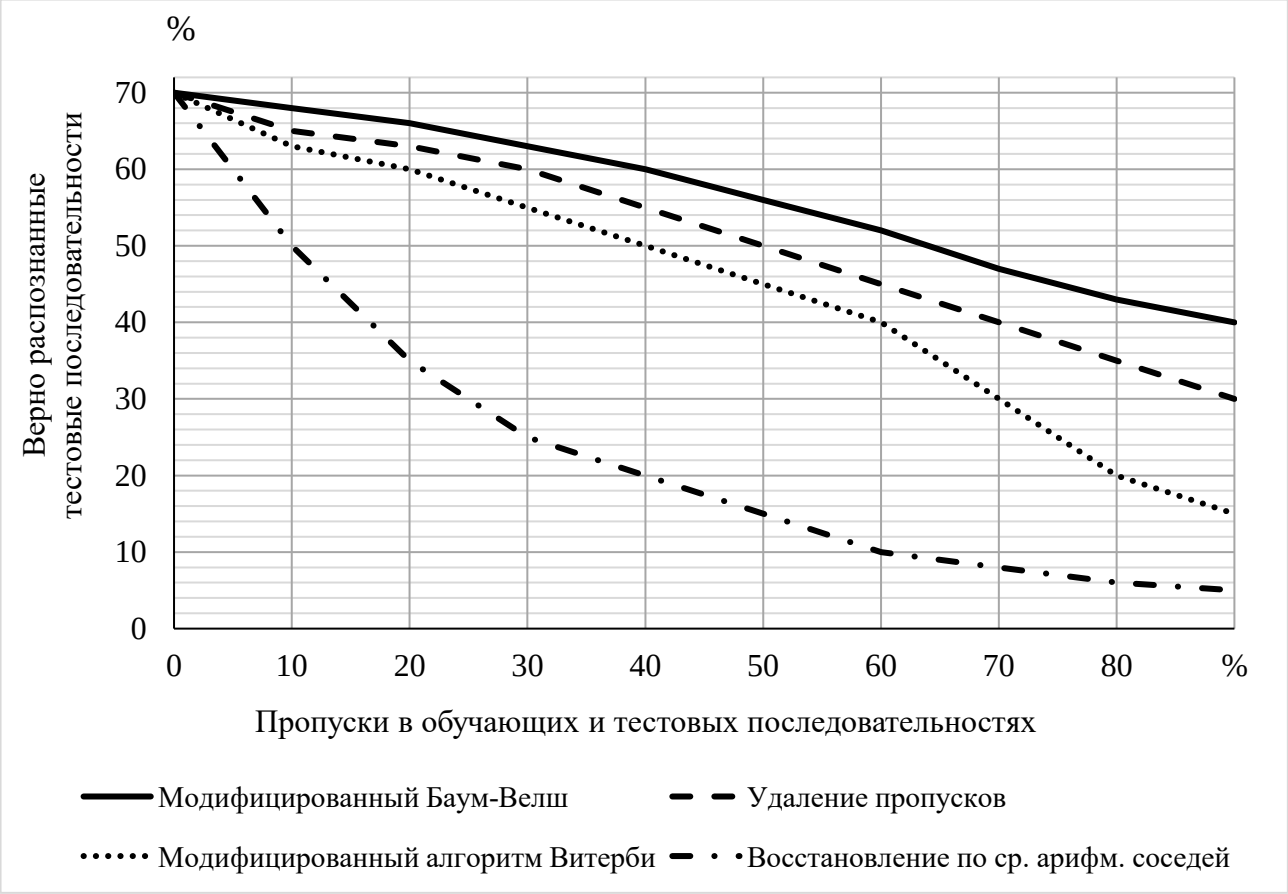

Рисунок 26 – Эффективность методики идентификации личности по неполным данным двигательной активности при неполной обучающей выборке

Как видно из рисунка 26, наилучший результат демонстрирует метод распознавания последовательностей на основе модифицированного алгоритма forward-backward, который использовал СММ, обученные с помощью модифицированного алгоритма Баума-Велша. Данный алгоритм позволяет до 1.3 раз увеличить точность идентификации пользователя по неполным данным его двигательной активности до 1.3 раз по сравнению с другими подходами с использованием СММ.

Выводы по четвертой главе

В данной главе был предложен и научно обоснован метод обучения скрытых марковских моделей по последовательностям с пропусками, основанный на модифицированном алгоритме Баума-Велша. Преимущество предложенного алгоритма по сравнению с ранее известными подходами было подтверждено экспериментально. Он оказался эффективнее, чем следующие подходы: предварительное условное восстановлением последовательностей с помощью модифицированного алгоритма Витерби; удаление пропущенных наблюдений из неполных последовательностей; предварительное восстановление пропусков по моде в случае дискретного распределения наблюдений или по среднему арифметическому соседей в случае непрерывного распределения наблюдений.

Разработанный метод также был успешно применен для решения задачи идентификации личности по неполным данным двигательной активности, при наличии неполных последовательностей, как в обучающей, так и в тестовой выборке. Таким образом, была подтверждена применимость разработанного метода к решению прикладной задачи, а также подтверждены результаты теоретических выкладок.

ГЛАВА 5 РАЗРАБОТКА МЕТОДА РАСПОЗНАВАНИЯ НЕПОЛНЫХ ПОСЛЕДОВАТЕЛЬНОСТЕЙ, ОПИСЫВАЕМЫХ СКРЫТЫМИ МАРКОВСКИМИ МОДЕЛЯМИ, БЛИЗКИМИ ПО ПАРАМЕТРАМ

В данной главе описан новый метод на основе модифицированного алгоритма вычисления первых производных от логарифма функции правдоподобия того, что случайный процесс, описываемый скрытой марковской моделью, сгенерировал неполную последовательность. Доказано, что данный метод можно применять для распознавания неполных последовательностей, описываемых скрытыми марковскими моделями, обученными на неполных последовательностях. Проведен сравнительный анализ эффективности разработанного метода и других подходов к распознаванию неполных последовательностей, описываемых скрытыми марковскими моделями.

5.1 Вычисление первых производных от функции правдоподобия того, что случайный процесс, описываемый скрытой марковской моделью, сгенерировал неполную последовательность

Рассмотрим способ вычисления производных от функции правдоподобия по параметрам СММ $\frac{\partial P\left(O|\lambda\right)}{\partial \eta}$, где η - некоторый параметр СММ, для СММ с непрерывной плотностью распределения наблюдений, при условии того, что последовательности, описываемые СММ, могут содержать пропуски. Используем приём маргинализации, описанный в разделе 2.1 для вывода формул модифицированного алгоритма вычисления первых производных.

Для сокращения записей введём следующие выражения: $b^*_i(\boldsymbol{o}_t)=\begin{cases} b_i(\boldsymbol{o}_t), & \boldsymbol{o}_t \neq \varnothing \\ 1, & \boldsymbol{o}_t = \varnothing \end{cases}$, $i=\overline{1,N}$, $t=\overline{1,T}$ и $g^*(\boldsymbol{o}_t;\mu_{im},\Sigma_{im})=\begin{cases} g(\boldsymbol{o}_t;\mu_{im},\Sigma_{im}), & \boldsymbol{o}_t \neq \varnothing \\ 1, & \boldsymbol{o}_t = \varnothing \end{cases}$, $i=\overline{1,N}$, $t=\overline{1,T}$, $m=\overline{1,M}$. Большая часть формул исходного алгоритма (41)-(58) останется без изменений, за исключением следующих:

формула (45) примет вид

$$\frac{\partial\overline{\alpha}_t(i)}{\partial\eta}=\left[\sum_{j=1}^{N}\left(\frac{\partial\hat{\alpha}_{t-1}(j)}{\partial\eta}a_{ji}+\hat{\alpha}_{t-1}(j)\frac{\partial a_{ji}}{\partial\eta}\right)\right]b^*{}_i(\boldsymbol{o}_t)+\sum_{j=1}^{N}\left(\hat{\alpha}_{t-1}(j)a_{ji}\right)\frac{\partial b^*{}_i(\boldsymbol{o}_t)}{\partial\eta},\ \begin{matrix}i=\overline{1,N}\\ t=\overline{2,T}\end{matrix};$$

формула (48) примет вид

$$\frac{\partial\overline{\alpha}_1(i)}{\partial\pi_j}=\begin{cases}b^*{}_i(\boldsymbol{o}_1),\ i=j\\ 0,\quad i\neq j\end{cases},\quad i,j=\overline{1,N};$$

формула (53) примет вид

$$\frac{\partial b_i(\boldsymbol{o}_t)}{\partial\tau_{i_1m}}=\begin{cases}g^*(\boldsymbol{o}_t;\mu_{im},\Sigma_{im}),\ i=i_1\\ 0,\qquad\qquad i\neq i_1\end{cases},\quad i,i_1=\overline{1,N},\ t=\overline{1,T},\ m=\overline{1,M};$$

формула (55) примет вид

$$\frac{\partial b_i(\boldsymbol{o}_t)}{\partial\mu^z_{i_1m}}=\begin{cases}0.5\tau_{im}g(\boldsymbol{o}_t;\mu_{im},\Sigma_{im})\dfrac{\boldsymbol{o}^z_t-\mu^z_{im}}{\Sigma^{zz}_{im}},\ i=i_1\\ 0,\qquad\qquad\qquad\qquad i\neq i_1\end{cases};$$

$$i,i_1=\overline{1,N},\ t=\overline{1,T},\ m=\overline{1,M},\quad z=\overline{1,Z}$$

формула (57) примет вид

$$\frac{\partial b_i(t)}{\partial\Sigma^{zz}_{i_1m}}=\begin{cases}0.5\tau_{im}g(\boldsymbol{o}_t;\mu_{im},\Sigma_{im})\left(\left(\dfrac{\boldsymbol{o}^z_t-\mu^z_{im}}{\Sigma^{zz}_{im}}\right)^2-\dfrac{1}{\left|\Sigma_{im}\right|}\right),\ i=i_1\\ 0,\qquad\qquad\qquad\qquad\qquad\qquad i\neq i_1\end{cases}.$$

$$i,i_1=\overline{1,N},\ t=\overline{1,T},\ m=\overline{1,M},\quad z=\overline{1,Z}$$

Полученный модифицированный вычисления производных от логарифма функции правдоподобия по параметрам СММ позволяет производить вычисления производных в том числе по неполным последовательностям [92].

5.2 Распознавание неполных последовательностей в пространстве первых производных от функции правдоподобия того, что случайный процесс, описываемый скрытой марковской моделью, сгенерировал неполную последовательность

Рассмотрим, каким образом можно проводить классификацию неполной последовательности по вычисленным на её основе производным логарифма правдоподобия того, что СММ сгенерировала данную последовательность. Идея состоит

в том, чтобы для каждой последовательности составить вектор признаков, состоящий из производных по параметрам всех конкурирующих СММ, а затем проводить классификацию полученных векторов одним из стандартных методов машинного обучения, например, методом опорных векторов [93, 94].

Опишем процедуру обучения описанного выше классификатора. Пусть дана обучающая выборка, т. е. множество неполных последовательностей наблюдений O^*, а также несколько СММ $\lambda_1, \lambda_2, \ldots, \lambda_D$ с одинаковым количеством параметров, причём для каждой неполной последовательности из обучающей выборки известно, какой их СММ она была порождена. Образуем для каждой неполной последовательности из обучающей выборки вектор признаков из D блоков, каждый из которых состоит из первых производных функции правдоподобия того, что данная последовательность была порождена СММ λ_d, по всем параметрам данной СММ λ_d, $d = \overline{1,D}$. Например, для неполной последовательности O данный вектор будет выглядеть следующим образом: $\left\{ \frac{\partial P\left(O|\lambda_1\right)}{\partial \eta_1}, \ldots, \frac{\partial P\left(O|\lambda_1\right)}{\partial \eta_L}, \ldots, \frac{\partial P\left(O|\lambda_D\right)}{\partial \eta_1}, \ldots, \frac{\partial P\left(O|\lambda_D\right)}{\partial \eta_L} \right\}$, где L – количество параметров СММ. Из полученного множества векторов сформируем выборку для обучения классификатора на основе метода опорных векторов [95]. Поскольку используемый классификатор является бинарным, то для рассматриваемого мультиклассового случая подойдёт стратегия обучения классификаторов one-vs-all (один против всех) [96-97]. Для подбора оптимальных гиперпараметров алгоритма рекомендуется использовать процедуру кросс-валидации [98].

После завершения обучения классификатора новую последовательность O можно классифицировать следующим образом. Для начала необходимо сформировать для этой последовательности вектор признаков описанным выше способом. Затем классифицируем полученный вектор с помощью построенных классификаторов, основанных на метода опорных векторов [99].

5.3 Оценка эффективности распознавания неполных последовательностей в пространстве первых производных от функции правдоподобия того, что случайный процесс, описываемый скрытой марковской моделью, сгенерировал неполную последовательность

Для оценки эффективности разработанный метод распознавания неполных последовательностей в пространстве первых производных от логарифма функции правдоподобия сравнивается с методом распознавания неполных последовательностей с помощью маргинализации пропущенных наблюдений (модифицированные алгоритмы Баума-Велша и forward-backward).

В качестве истинных СММ были взяты модели λ_1 и λ_2 со следующими характеристиками. Число скрытых состояний $N=3$, количество компонент в смесях $M=3$. Размерность векторов наблюдений $Z=2$. Вектор распределения начального состояния: $\Pi=[1,0,0]$, матрица вероятностей переходов: $A=\begin{bmatrix} 0.1+\Delta A & 0.7-\Delta A & 0.2 \\ 0.2 & 0.2+\Delta A & 0.6-\Delta A \\ 0.8-\Delta A & 0.1 & 0.1+\Delta A \end{bmatrix}$, веса компонент смесей $\left\{\tau_{im}, i=\overline{1,N}, m=\overline{1,M}\right\}=\begin{pmatrix} 0.3+\Delta\tau & 0.4-\Delta\tau & 0.3 \\ 0.3 & 0.4+\Delta\tau & 0.3-\Delta\tau \\ 0.3-\Delta\tau & 0.4 & 0.3+\Delta\tau \end{pmatrix}$ (номеру строки соответствует номер скрытого состояния, а номеру столбца – номер компоненты смеси), вектора математических ожиданий компонент смесей $\left\{\mu_{im}, i=\overline{1,N}, m=\overline{1,M}\right\}=\begin{pmatrix} (0-\Delta\mu \quad 0+\Delta\mu)^T & (1-\Delta\mu \quad 1+\Delta\mu)^T & (2-\Delta\mu \quad 2+\Delta\mu)^T \\ (3-\Delta\mu \quad 3+\Delta\mu)^T & (4-\Delta\mu \quad 4+\Delta\mu)^T & (5-\Delta\mu \quad 5+\Delta\mu)^T \\ (6-\Delta\mu \quad 6+\Delta\mu)^T & (7-\Delta\mu \quad 7+\Delta\mu)^T & (8-\Delta\mu \quad 8+\Delta\mu)^T \end{pmatrix}$ (номеру строки соответствует номер скрытого состояния, а номеру столбца – номер компоненты смеси), все ковариационные матрицы компонент смесей $\left\{\Sigma_{im}, i=\overline{1,N}, m=\overline{1,M}\right\}$ были выбраны диагональными, значение всех элементов на

диагонали было равно $0.1+\Delta\sigma$. При этом у первой модели $\Delta A=0$, $\Delta\tau=0$, $\Delta\mu=0$, $\Delta\sigma=0$, а у второй модели $\Delta A=0.05$, $\Delta\tau=0.05$, $\Delta\mu=0.01$, $\Delta\sigma=0.01$. Такой выбор параметров максимально усложняет задачу распознавания, поскольку случайные процессы, описываемые такими моделями, очень близки по свойствам и порождаемые ими последовательности трудно различить. С помощью каждой из моделей λ_1 и λ_2 было сгенерировано $K=100$ обучающих и тестовых последовательностей длиной $T=100$, причем каждая из последовательностей содержала G пропусков (число G изменялось от 0 до 90 в ходе эксперимента) в случайных местах. С помощью обучающих неполных последовательностей были получены оценки моделей λ_1 и λ_2 по алгоритму обучения СММ по неполным обучающим последовательностям, основанном на маргинализации пропущенных наблюдений. Также с помощью производных от обучающих последовательностей и оценок СММ был обучен классификатор метода опорных векторов, гиперпараметры которого были подобраны с помощью кросс-валидации по четырём блокам. Затем с помощью λ_1 и λ_2 проводилось распознавание неполных тестовых последовательностей с помощью модифицированного алгоритма forward-backward по критерию максимума функции правдоподобия (штриховая линия) и с помощью первых производных от логарифма функции правдоподобия, используя метод опорных векторов в качестве классификатора (сплошная линия), причем использовались производные по всем параметрам моделей. Фиксировалось количество правильно распознанных последовательностей. Рисунок 27 содержит усреднённые результаты после 100 запусков описанного выше эксперимента с различными начальными значениями генератора случайных чисел.

Как видно, метод распознавания, основанный на производных, начинает превосходить метод, основанный на маргинализации пропущенных наблюдений (модифицированные алгоритмы Баума-Велша и forward-backward), начиная примерно с 20% пропусков в обучающих и тестовых последовательностях.

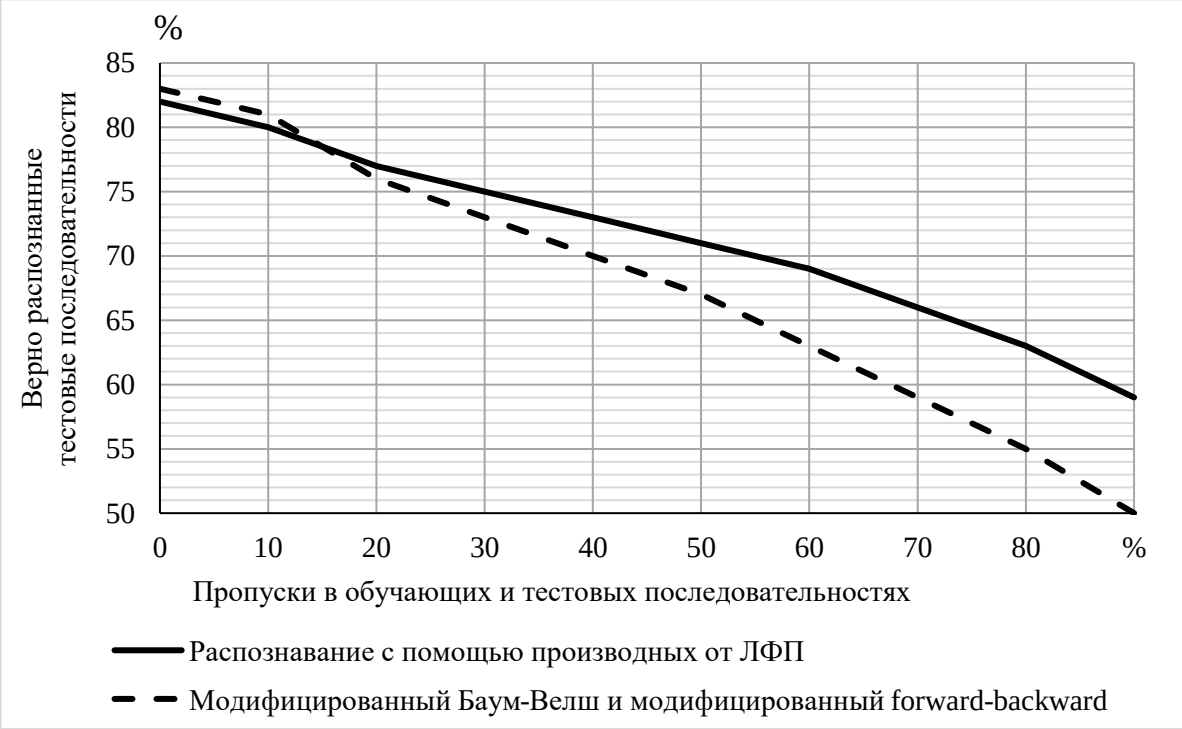

Рисунок 27 – Оценка эффективности метода распознавания неполных последовательностей в пространстве первых производных от логарифма функции правдоподобия

При этом преимущество метода на основе производных увеличивается с увеличением процента пропусков, достигая 1.2 раза при 90% пропусков в последовательностях [100].

5.4 Разработка методики идентификации личности по неполным данным двигательной активности с использованием производных от логарифма функции правдоподобия

Данная практическая задача заключается в обучении СММ на неполных последовательностях, генерируемых носимым устройством, анализирующим данные двигательной активности, и дальнейшего распознавания неполных последовательностей. Поскольку двигательная активность многих людей достаточно схожа, алгоритм классификации, основанный на производных, потенциально может увеличить точность идентификации.

В качестве исходных данных использовалась та же выборка двигательной активности, что и в п. 2.7.1. Для каждого участника эксперимента по соответствующим ему последовательностям была обучена своя СММ с помощью модифицированного алгоритма Баума-Велша (п. 4.1). Далее, с помощью процедуры, описанной в п. 5.2, был обучен набор классификаторов метода опорных векторов, причём гиперпараметры классификаторов были получены с помощью кросс-валидации по 4 блокам на обучающей выборке. Каждая СММ имела N=3 скрытых состояния и M=3 компонент смесей из 3-хмерных (Z=3) нормальных распределений. Число скрытых состояний и компонент смесей были подобраны эмпирически таким образом, чтобы обеспечить наибольшую точность при приемлемом времени расчёта. Каждая из больших последовательностей была разделена на подпоследовательности длиной T=100 (что соответствует примерно 3-м секундам наблюдения). Для обучения и кросс-валидации было использовано 75% случайно выбранных последовательностей из каждого класса. Тестирование обученных моделей проводилось на 25% оставшихся последовательностей. Для оценки качества классификации измерялся средний процент правильно распознанных последовательностей из каждого класса. Вышеупомянутая метрика была рассчитана для разного количества пропусков в обучающих и тестовых последовательностях, а также для двух различных алгоритмов распознавания неполных последовательностей, указанных далее. Расположение пропусков было выбрано случайно в каждой из обучающих и распознаваемых последовательностей.

Рисунок 28 содержит результаты данного эксперимента. На графике приведены средние значения после 100 проведений эксперимента с различным начальным значением генератора случайных чисел. Тип линии обозначает использованный метод распознавания: сплошная – на основе производных (п. 3.1), а штриховая – на основе модифицированного алгоритма forward-backward (п. 5.2) [84].

Как видно из рисунка, метод распознавания, основанный на производных, превосходит метод, основанный на маргинализации пропущенных наблюдений (модифицированные алгоритмы Баума-Велша и forward-backward).

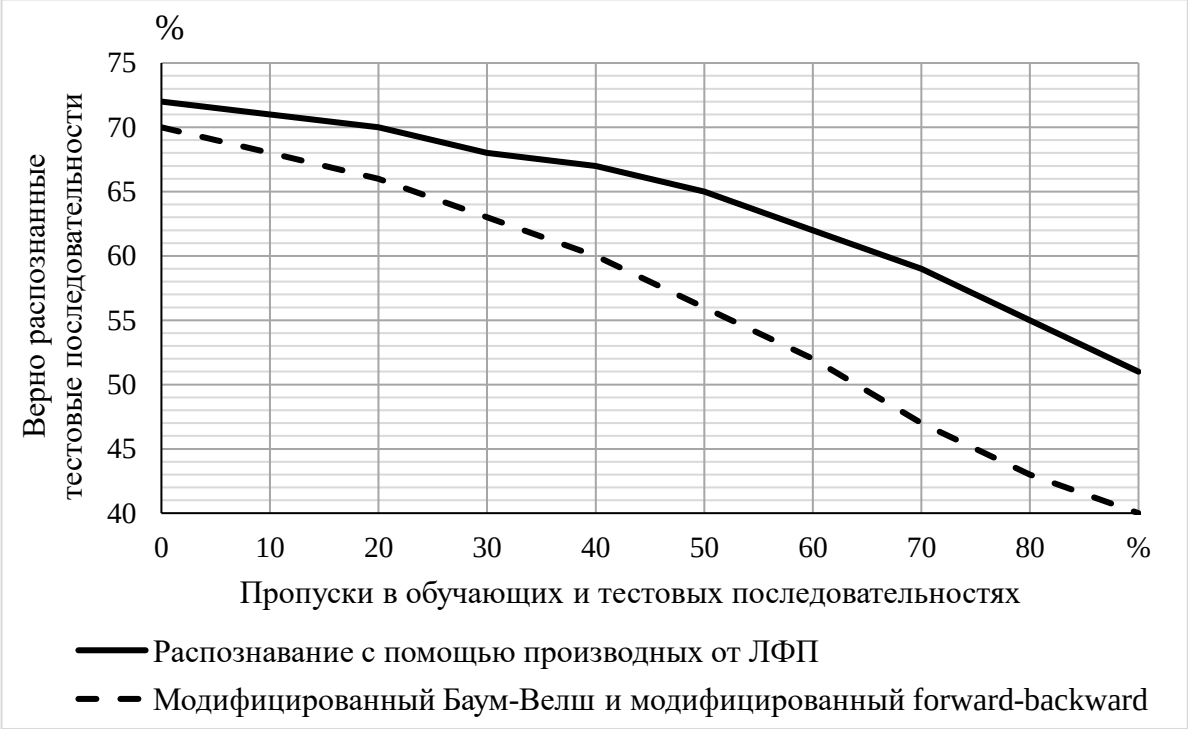

Рисунок 28 – Оценка эффективности методики распознавания личности по неполным данным с помощью производных от логарифма функции правдоподобия

Причём при увеличении процента пропусков в обучающих и тестовых последовательностях его точность также увеличивается, становясь на 12% выше, чем у других алгоритмов, при 90% пропусков в последовательностях.

Выводы по пятой главе

В данной главе был предложен и научно обоснован метод распознавания неполных последовательностей, который заключается в классификации последовательностей в пространстве признаков, образованном первыми производными от логарифма функции правдоподобия того, что случайный процесс, описываемый скрытой марковской моделью, сгенерировал распознаваемую неполную последовательность. Сравнительный анализ предложенного метода и разработанного автором ранее метода распознавания неполных последовательностей, основанного на модифицированном алгоритме forward-backward, показал, что предложенный ме-

тод на основе производных позволяет достичь большего количества правильно распознанных последовательностей, чем метод на основе критерия максимума функции правдоподобия, начиная с некоторого (в проведенном эксперименте – более 20%) процента пропусков в обучающих и тестовых последовательностях. Таким образом, предложенный метод может быть рекомендован к применению в условиях сильных помех, когда имеется много пропущенных данных, однако распознавание неполных последовательностей всё же необходимо проводить.

Разработанный метод также был успешно применен для решения задачи идентификации личности по неполным данным двигательной активности, при наличии неполных последовательностей, как в обучающей, так и в тестовой выборке. Метод распознавания последовательностей, основанный на производных от логарифма функции правдоподобия по параметрам СММ, показал наилучший результат по сравнению с алгоритмом распознавания, основанным на модифицированном алгоритме forward-backward, особенно при большом проценте пропусков. Таким образом, была доказана применимость разработанных методов к решению прикладной задачи, а также подтверждены теоретические выкладки и результаты синтетических экспериментов.

ЗАКЛЮЧЕНИЕ

В диссертационной работе на основании разработки и исследования методов анализа неполных последовательностей данных решена задача нивелирования пропусков путём применения аппарата скрытых марковских моделей и приема маргинализации пропущенных наблюдений, что имеет важное значение для развития теории скрытых марковских моделей и решения практических задач.

1) Разработан и исследован метод декодирования и восстановление неполных последовательностей, который обеспечивает:

декодирование состояний скрытой марковской модели до 1.3 раза точнее при дискретном распределении наблюдений и до 1.4 раза точнее при непрерывном распределении наблюдений, чем при использовании стандартных методов СММ;

восстановление пропусков в неполных последовательностях до 1.2 раз точнее при дискретном распределении наблюдений и до 7 раз точнее при непрерывном распределении наблюдений, чем при использовании стандартных методов СММ.

2) Разработан и исследован метод распознавания неполных последовательностей, который позволяет проводить классификацию неполных последовательностей до 1.3 раз точнее при дискретном распределении наблюдений и до 1.6 раз точнее при непрерывном распределении наблюдений, чем при использовании стандартных методов СММ.

3) Разработан и исследован метод обучения скрытых марковских моделей по неполным последовательностям, который позволяет увеличить до 1.2 раз точность распознавания последовательностей с помощью обученных моделей при дискретном и непрерывном распределении наблюдений, чем при использовании стандартных методов обучения и распознавания с помощью СММ.

4) Разработан и исследован метод распознавания неполных последовательностей, основанный на алгоритме вычисления первых производных от логарифма функции правдоподобия того, что скрытая марковская модель породила неполную последовательность. Метод позволяет увеличить до 1.2 раза точность распознава-

ния подобных последовательностей по сравнению с методом распознавания неполных последовательностей, основанным на модифицированном алгоритме forward-backward.

5) На основании теоретических исследований решены три практические задачи и разработаны методики:

— декодирования маршрута абонента по транспортному графу, который соответствует последовательности его регистраций в мобильной сети. Она позволяет вычислить траекторию абонента в 2.5 раза точнее, чем при использовании существующего метода соединения центроид покрытий секторов последовательных регистраций;

— восстановления неполных данных двигательной активности человека, превосходящая по точности стандартный метод CMM до 5 раз;

— идентификации личности по неполным данным двигательной активности, позволяющая повысить точность идентификации до 1.3 раза по сравнению со стандартным методом CMM, предполагающим предварительное исключение пропусков из последовательностей.

Дальнейшие исследования предполагают апробацию и адаптацию разработанных методов для различных задач анализа неполных данных.

СПИСОК СОКРАЩЕНИЙ

СММ – скрытая марковская модель;

HMM – hidden Markov model (скрытая марковская модель);

GMM – gaussian mixture model (модель смесей нормальный распределений);

ЛФП – логарифм функции правдоподобия;

МФП – максимум функции правдоподобия;

EM – expectation-maximization (ожидание-максимизация);

KNN – k-nearest neighbors (K-ближайших соседей);

SVM – support vector machines (метод опорных векторов);

БС – базовая станция;

GSM – Global System for Mobile communication (глобальная система мобильной коммуникации);

GPS – global positioning system (глобальная система позиционирования);

LTE – long-term evolution (долговременная эволюция);

2g, 3g, 4g – 2-generation, 3-generation, 4-generation (generation - поколение);

MAD – median absolute deviation (медианное абсолютное отклонение);

п. – пункт.

СПИСОК УСЛОВНЫХ ОБОЗНАЧЕНИЙ

t – индекс момента времени (номер наблюдения в последовательности);

T – длина последовательности наблюдений;

s_i – скрытое состояние СММ под номером i;

q – скрытое состояние без привязки к моменту времени;

q_t – скрытое состояние в момент времени t;

o – наблюдение без привязки к моменту времени;

o_t – наблюдение в момент времени t;

N – количество скрытых состояний СММ;

M – размер алфавита наблюдений СММ, либо количество компонент в смеси многомерных нормальных распределений, описывающих распределение наблюдения, порождаемые СММ;

λ – СММ (конкретный набор параметров СММ);

Π – дискретное распределение начального скрытого состояния СММ;

π_i – вероятность того, что начальное скрытое состояние СММ имеет номер i;

A – вероятности перехода из одного скрытого состояния СММ в другое;

A_{ij} – вероятность перехода СММ из скрытого состояния с номером i в скрытое состояние с номером j;

B – распределение наблюдений при условии нахождении СММ в определённом скрытом состоянии;

$b_i(o)$ – вероятность породить наблюдение o при нахождении в скрытом состоянии i;

V – конечный алфавит наблюдений СММ с дискретным распределением наблюдений;

v_m – символ алфавита СММ под номером m;

$\mathbb{R}$ – множество действительных чисел;

Z – размерность наблюдений СММ с непрерывным распределением наблюдений;

τ_{im} – вес m-го компонента смеси нормальных распределений при нахождении СММ в i-м состоянии;

μ_{im} – математическое ожидание m-го компонента смеси при нахождении СММ в i-м состоянии;

Σ_{im} – ковариационная матрица m-го компонента смеси при нахождении СММ в i-м состоянии;

$g(...)$ – функция плотности многомерного нормального распределения;

$\hat{Q}$ – наиболее вероятная последовательность скрытых состояний СММ по алгоритму Витерби;

$\hat{q}_t$ – наиболее вероятное скрытое состояние СММ по алгоритму Витерби в момент времени t;

$\delta_t(i)$ – максимальная вероятность того, что СММ находится в скрытом состоянии под номером i в момент времени t с учётом всех предыдущих наблюдений;

$\psi_t(i)$ – номер предыдущего скрытого состояния, при котором достигается максимальная вероятность того, что СММ находится в скрытом состоянии под номером i в момент времени t с учётом всех предыдущих наблюдений;

L – значение функции правдоподобия того, что СММ породила некоторую последовательность;

$\alpha_t(i)$ – прямая вероятность или вероятность того, что СММ находится в состоянии i в момент времени t с учётом всех предыдущих наблюдений;

$\beta_t(i)$ – обратная вероятность или вероятность того, что СММ находится в состоянии i в момент времени t с учётом всех последующих наблюдений;

$\gamma_t(i)$ – вероятность того, что СММ находится в состоянии под номером i в момент времени t с учётом всей последовательности наблюдений;

$\xi_t(i,j)$ – вероятность того, что СММ находится в состоянии под номером i в момент времени t и переходит в состояние под номером j в следующий момент времени с учётом всей последовательности наблюдений;

$\gamma_t(i,m)$ – вероятность того, что СММ находится в состоянии под номером i в момент времени t и при этом для генерации наблюдения была использована компонента смеси распределений под номером m с учётом всей последовательности наблюдений;

ω_{it} - компонента смеси распределений в момент времени t в скрытом состоянии СММ под номером i;

K – количество последовательностей наблюдений;

$\hat{x}$ – оценка x, где x – один из параметров СММ;

$\hat{x}'$ – новое приближение оценки x, где x – один из параметров СММ;

$\tilde{M}$ – количество компонент смеси в модели GMM;

$\tilde{\tau}_m$ – вес m-го компонента смеси в модели GMM;

$\tilde{\mu}_m$ – математическое ожидание m-го компонента в модели GMM;

$\tilde{\Sigma}_m$ – ковариационная матрица m-го компонента в модели GMM;

θ – модель GMM (конкретный набор параметров модели GMM);

θ_m – параметры GMM для m-й компоненты смеси;

$\tilde{\delta}_t(i)$ – отмасштабированная версия $\delta_t(i)$;

$\tilde{\psi}_t(i)$ – отмасштабированная версия $\psi_t(i)$;

C_t – параметр масштаба для прямой вероятности в момент времени t;

c_t – вспомогательный параметр масштаба для прямой вероятности в момент времени t;

$\hat{\alpha}_t(i)$ – отмасштабированная прямая вероятность;

$\bar{\alpha}_t(i)$ – вспомогательная отмасштабированная прямая вероятность;

D_t – параметр масштаба для обратной вероятности в момент времени t;

$\hat{\beta}_t(i)$ – отмасштабированная обратная вероятность;

$\bar{\beta}_t(i)$ – промежуточная отмасштабированная обратная вероятность;

η – некоторый параметр СММ, по которому вычисляется производная логарифма функции правдоподобия того, что СММ породила последовательность;

q_t^* – состояние, в котором находилась СММ в момент времени t при искусственной генерации последовательностей;

$\varnothing$ – пропущенное наблюдение;

V^* – область определения наблюдений СММ с дискретным распределением, допускающая пропуск;

R^* – область определения наблюдений СММ с непрерывным распределением, допускающая пропуски;

G – количество пропусков в последовательности;

Δx – смещение x, где x – один из параметров СММ.

СЛОВАРЬ ТЕРМИНОВ

Маргинализация – приём использования маргинального распределения, т. е. распределения некоторых случайных величин без указания на значения других случайных величин;

Неполная последовательность наблюдений – последовательность наблюдений, где значение некоторых наблюдений не определено, причем предполагается, что пропуски возникают в случайных местах последовательности без какой-либо закономерности;

Склеивание неполных последовательностей – метод, подразумевающий удаления пропусков из неполных последовательностей с целью их дальнейшего анализа с помощью стандартных алгоритмов;

Декодирование последовательности наблюдений, описываемой СММ – нахождение оптимальной в некотором смысле последовательности скрытых состояний, соответствующей последовательности наблюдений;

Распознавание (классификация) объекта – определение класса, к которому принадлежит объект;

Map matching (сопоставление с картой) – вычисление наиболее вероятного маршрута по графу на основе последовательности примерных пространственных координат;

Базовая станция – техническое сооружение, на которое устанавливаются антенны и другое телекоммуникационное оборудование, позволяющее обслуживать мобильную сеть;

Сектор мобильной сети – условный участок земли, покрываемой одной из антенн, установленных на базовой станции;

Зона местоположения – группа секторов мобильной сети, позволяющая установить примерное место нахождения устройства;

Регистрация абонента в мобильной сети – любое событие (например, звонок, СМС-сообщение, передача интернет-трафика, смена зоны местоположения) в мобильной сети, связанное с активностью устройства абонента, позволяющее установить сектор мобильной сети, где произошло событие, а также время наступления события.

СПИСОК ЛИТЕРАТУРЫ

1. Baum, L. E. A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains / L. E. Baum [et. al.] // The Annals of Mathematical Statistics. – 1970. – Vol. 41, №1. – pp. 164-171.
2. Baum, L. E. Statistical inference for probabilistic functions of finite state Markov chains / L. E. Baum, T. Petrie // The Annals of Mathematical Statistics. – 1966. – Vol. 37. – pp. 1554-1563.
3. Baum, L. E. An inequality with applications to statistical estimation for probabilistic functions of a Markov process and to a model for ecology / L. E. Baum, J.A. Eagon // Bulletin of the American Mathematical Society. – 1967. – Vol. 73, № 3. – pp. 360-363.
4. Baum, L. E. Growth functions for transformations / L. E. Baum, G. R. Sell // Pacific journal of Mathematics. – 1968. – Vol. 27, №2. – pp. 211-227.
5. Baum, L. E. An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes / L. E. Baum // Inequalities. – 1972. – Vol. 3. – pp. 1-8.
6. Gales, M. The Application of Hidden Markov Models in Speech Recognition / M. Gales, S. Young // Signal Processing. – 2007. – Vol. 1, №3. – pp. 195-304.
7. Levinson, S. E. An introduction to the application of the theory of probabilistic functions of a Markov process to automatic speech recognition / S. E. Levinson, L. R. Rabiner, M. M. Sondhi // The Bell System Technical Journal. – 1983. – Vol. 62, №4. – pp. 1035-1074.
8. Baker, J. K. Some Experiments in Automatic Recognition of Continuous Speech / J. K. Baker, L. R. Bahl // Proceedings of the 11th Annual IEEE Computer Society Conference. – 1975. – pp. 326-329.
9. Munteanu, D. P. Automatic speaker verification experiments using HMM / D. P. Munteanu, S. A. Toma // International Conference on Communications (COMM). – 2010. – pp. 107-110.
10. Gales, M. J. Maximum likelihood linear transformations for HMM-based speech recognition / M. J. Gales // Computer Speech & Language. – 1998. – Vol. 12, №2. – pp. 75-98.

11. Birney, E. Hidden Markov models in biological sequence analyzing / E. Birney // IBM Journal Research & Development. – 2001. – Vol. 45, № 3/4. – pp. 449-449.

12. Bishop, M. J. Maximum Likelihood Alignment of DNA Sequences / M. J. Bishop, E. A. Thompson // Journal of Molecular Biology. – 1986. – Vol. 190, № 2. – pp. 159-165.

13. Söding, J. Protein homology detection by HMM–HMM comparison / J. Söding // Bioinformatics. – 2005. – Vol. 21, №7. – pp. 951–960.

14. Käll, L. An HMM posterior decoder for sequence feature prediction that includes homology information / L. Käll, A. Krogh, E. L. Sonnhammer // Bioinformatics. – 2005. – Vol. 21, №1. – pp. i251–i257.

15. Malekpour, S. A. MGP-HMM: Detecting genome-wide CNVs using an HMM for modeling mate pair insertion sizes and read counts / S. A. Malekpour, H. Pezeshk, M. Sadeghi // Mathematical Biosciences. – 2016. – Vol. 279. – pp. 53-62.

16. Bhar, R. Hidden Markov Models: Applications to Financial Economics (Advanced Studies in Theoretical and Applied Econometrics) / R. Bhar, S. Hamori. – London, New York: Springer, 2004. – 160 p.

17. Erlwein, C. An examination of HMM-based investment strategies for asset allocation / C. Erlwein, R. Mamon, M. Davison // Applied Stochastic Models in Business and Industry. – 2011. – Vol. 27, №3. – pp. 204-221.

18. McCulloch, R. E. Statistical analysis of economic time series via Markov switching models / R. E. McCulloch, R. S. Tsay // Journal of Time Series Analysis. – 1994. – Vol. 15, №5. – pp. 523-539.

19. Rossia, A. Volatility estimation via hidden Markov models / A. Rossia, G. M. Gallob // Journal of Empirical Finance. – 2006. – Vol.13, №2. – pp. 203-230.

20. Bunke, H. Hidden Markov models : applications in computer vision / H. Bunke, T. Caelli. – Singapore: WorldScientific, 2001. – 244 p.

21. Nefian, A. V. An embedded HMM-based approach for face detection and recognition / A. V. Nefian, M. H. Hayes // Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing. – 1999. – Vol. 6. – pp. 3553-3556.

22. Niu, F. HMM-Based Segmentation and Recognition of Human Activities from Video Sequences / F. Niu, M. Abdel-Mottaleb // IEEE International Conference on Multimedia and Expo. – 2005. – pp. 804-807.

23. Zhang, L. A Vision-based Sign Language Recognition System Using Tied-mixture Density HMM / L. Zhang, Y. Chen, G. Fang, X. Chen, W. Gao // Proceedings of the 6th International Conference on Multimodal Interfaces. – 2004. – pp. 198-204.

24. Rabiner, L. R. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition / L. R. Rabiner // Proceedings of the IEEE. – 1989. – Vol. 77. – pp. 257-285.

25. Статистика упоминания ключевого слова “hidden Markov models” между 1800 и 2008 годами, полученная с помощью сервиса Google Ngram Viewer [Электронный ресурс]. – Режим доступа: https://books.google.com/ngrams/graph?content=hidden+Markov+models&year_start=1800&year_end=2008&corpus=15&smoothing=3&share=&direct_url=t1%3B%2Chidden%20Markov%20models%3B%2Cc0

26. Altun, K. Comparative study on classifying human activities with miniature inertial and magnetic sensors / K. Altun, B. Barshan, O. Tunçel // Pattern Recognition. – 2010. – Vol. 43, №10. – pp. 3605-3620.

27. Barshan, B. Recognizing daily and sports activities in two open source machine learning environments using body-worn sensor units / B. Barshan, M. C. Yüksek // The Computer Journal. – 2014. – Vol. 57, №11. – pp. 1649-1667.

28. Altun, K. Human activity recognition using inertial/magnetic sensor units / K. Altun, B. Barshan // Proceedings First International Workshop on Human Behavior Understanding (in conjunction with the 20th Int. Conf. on Pattern Recognition). – 2010. – Vol. LNCS 6219. – pp. 38-51.

29. Casale, P. Personalization and user verification in wearable systems using biometric walking patterns / P. Casale, O. Pujol, P. Radeva // Personal and Ubiquitous Computing. – 2012. – Vol. 16, №5. – pp. 563-580.

30. Nickel, C. Classifying accelerometer data via hidden Markov models to authenticate people by the way they walk / C. Nickel, C. Busch // IEEE Aerospace and Electronic Systems Magazine. – 2013. – Vol. 28, №10. – pp. 29-35.

31. Nickel, C. Benchmarking the performance of SVMs and HMMs for accelerometer-based biometric gait recognition / C. Nickel, H. Brandt, C. Busch

// IEEE International Symposium on Signal Processing and Information Technology. – 2011. – pp. 281-286.

32. Newson, P. Hidden Markov Map Matching Through Noise and Sparseness / P. Newson, J. Krumm // Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems. – 2009. – pp. 336-343.

33. Koller, H. Fast Hidden Markov Model Map-Matching for Sparse and Noisy Trajectories / H. Koller, P. Widhalm, M. Dragaschnig, A. Graser // IEEE 18th International Conference on Intelligent Transportation Systems. – 2015. – pp. 2557-2561.

34. Goh, C. Y. Online map-matching based on Hidden Markov model for real-time traffic sensing applications / C. Y. Goh, J. Dauwels, N. Mitrovic, M. T. Asif, A. Oran, P. Jaillet // 15th International IEEE Conference on Intelligent Transportation Systems. – 2012. – pp. 776-781.

35. Mohamed, R. Accurate and Efficient Map Matching for Challenging Environments / R. Mohamed, H. Aly, M. Youssef // Proceedings of the 22Nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems. – 2014. – pp. 401-404.

36. Wang, G. Eddy: An Error-bounded Delay-bounded Real-time Map Matching Algorithm Using HMM and Online Viterbi Decoder / G. Wang, R. Zimmermann // Proceedings of the 22Nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems. – 2014. – pp. 33-42.

37. Cooke, M. Robust automatic speech recognition with missing and unreliable acoustic data / M. Cooke, P. Green, L. Josifovski, A. Vizinh // Speech Communication. – 2001. – Vol. 34, №3. – pp. 267-285.

38. Lee, D. Missing motion data recovery using factorial hidden Markov models / D. Lee, D. Kulic, Y. Nakamura // IEEE International Conference on Robotics and Automation. – 2008. – pp. 1722-1728.

39. Nielsen, J. Algorithms for a Parallel Implementation of Hidden Markov Models with a Small State Space / J. Nielsen, A. Sand // IEEE International Symposium on Parallel and Distributed Processing Workshops and Phd Forum (IPDPSW). – 2011. – pp. 452-459.

40. Jun, L The fast evaluation of hidden Markov models on GPU / L. Jun, C. Shuangping, L. Yanhui // ICIS 2009: IEEE International Conference on Intelligent Computing and Intelligent Systems. – 2009. – pp. 426-430.

41. Liu, C. cuHMM: a CUDA Implementation of Hidden Markov Model: Training and Classication [Электронный ресурс] / C. Liu. – 2009. – Режим доступа: https://liuchuan.org/pub/cuHMM.pdf

42. Lee, S. Final Project: Parallel Viterbi on a GPU [Электронный ресурс] / S. Lee. – Режим доступа: http://homes.cs.washington.edu/~miro/docs/HMM_on_GPU.pdf

43. Уваров, В. Е. Использование графических процессоров для оптимизации процесса классификации многомерных последовательностей, описываемых скрытыми марковскими моделями / В. Е. Уваров, В. А. Уварова, А. И. Фомин // Вестник Кузбасского Государственного Технического Университета. – 2015. – №1. – С. 88-91.

44. Gultyaeva, T. A. Graphics processing unit implementation of Hidden Markov models / T. A. Gultyaeva, A. S. Sautin, V. E. Uvarov // International Conference on Actual Problems of Electronic Instrument Engineering Proceedings (APEIE-2014). – 2014. – Vol. 1. – pp. 571-573.

45. Гультяева, Т. А. Использование графических процессоров для оптимизации работы скрытых марковских моделей / Т. А. Гультяева, А. С. Саутин, В. Е. Уваров // Труды XII международной конференции Актуальные проблемы электронного приборостроения (АПЭП-2014). – 2014. – Т. 6. – С. 83-86.

46. Гультяева, Т. А. Решение на GPU задачи обучения и классификации многомерных числовых последовательностей с помощью скрытых марковских моделей / Т. А. Гультяева, В. Е. Уваров // Материалы 53-й Международной научной студенческой конференции МНСК-2015. – 2015. – С. 196-196.

47. Гультяева, Т. А. Решение на GPU задач обучения скрытых Марковских моделей и распознания многомерных числовых последовательностей с их помощью / Т. А. Гультяева, В. Е. Уваров // Материалы российской научно-технической конференция "Обработка информации и математическое моделирование". – 2015. – С. 79-89.

48. Гультяева, Т. А. Применение гибридных вычислений для ускорения процесса машинного обучения с помощью скрытых марковских моделей /

Т. А. Гультяева, В. Е. Уваров // Сборник научных трудов конференции "Наука. Технологии. Инновации." – 2015. – С. 117-118.

49. Гультяева, Т. А. Использование гибридных вычислений для оптимизации процесса распознавания последовательностей, описываемых скрытыми марковскими моделями / Т. А. Гультяева, А. А. Попов, В. Е. Уваров // Сборник научных трудов НГТУ. – 2015. – №4. – С. 42-55.

50. Гультяева, Т. А. Optimization of multidimensional sequence recognition method based on hidden markov models using parallel hybrid computations / Т. А. Гультяева, А. А. Попов, В. Е. Уваров // Материалы 54-й международной научной студенческой конференции МНСК-2016. – 2016. – С. 138-138.

51. Khreich, W. A survey of techniques for incremental learning of HMM parameters / W. Khreich, E. Granger, A. Miri, R. Sabourin // Inf. Sci. – 2012. – Vol. 197. – pp. 105-130.

52. Florez-Larrahondo, G. Incremental estimation of discrete hidden Markov models based on a new backward procedure / G. Florez-Larrahondo, S. Bridges, E. A. Hansen // Proceedings of the 20th national conference on Artificial intelligence. – 2005. – Vol. 2. – pp. 758-763.

53. Cavalin, P. R. Evaluation of incremental learning algorithms for HMM in the recognition of alphanumeric characters / P. R. Cavalin, R. Sabourin, C. Y. Suen, A. S. Britto // Pattern Recognition. – 2009. – Vol. 42, №12. – pp. 3241-3253.

54. Chatzis, S. A Robust to Outliers Hidden Markov Model with Application in Text-Dependent Speaker Identification / S. Chatzis, T. Varvarigou // IEEE International Conference on Signal Processing and Communications. – 2007. – pp. 804-807.

55. Popov, A. A. The classification of noisy sequences generated by similar hmms / A. A. Popov, T. A. Gultyaeva // Lecture Notes in Computer Science. – 2011. – Vol. 6744. – pp. 30-35.

56. Гультяева, Т. А. Классификация зашумленных последовательностей, порожденных близкими скрытыми марковскими моделями / Т. А. Гультяева, А. А. Попов // Научный вестник НГТУ. – 2011. – №3. – С. 3-16.

57. Гультяева, Т. А. Классификация последовательностей, подверженных действию помех с характеристиками, зависящими от скрытых состояний /

Т. А. Гультяева, А. А. Попов // Сборник Научных трудов НГТУ. – 2011. – №1. – С. 59-68.

58. Гультяева, Т. А. Классификация последовательностей, порожденных близкими скрытыми марковскими моделями, при наличии шума / Т. А. Гультяева, А. А. Попов // Технические науки: проблемы и перспективы. – 2011. – №1. – С. 37-41.

59. Гультяева, Т. А. Классификация последовательностей, порожденных близкими скрытыми марковскими моделями, при наличии шума, распределенного по закону Коши / Т. А. Гультяева, А. А. Попов // Материалы российской науч.-технич. конф. Информатика и проблема телекоммуникаций. – 2011. – Т. I. – С. 60-63.

60. Bilmes, J. A. A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov models / J. A. Bilmes // Technical Report 97-021. – 1998.

61. Viterbi, A. J. Error Bounds for Convolutional Codes and an Asymptotically Optimum Decoding Algorithm / A. J. Viterbi // IEEE Transactions on Information Theory. – 1967. – №13. – pp. 260-269.

62. Шлезингер, М. Десять лекций по статистическому и структурному распознаванию / М. Шлезингер, В. Главач. – Киев: Наукова думка, 2004.

63. Collins, M. The Forward-Backward Algorithm [Электронный ресурс] / M. Collins. – Режим доступа: http://www.cs.columbia.edu/~mcollins/fb.pdf

64. Dempster, A. P. Maximum likelihood from incomplete data via the EM algorithm / A. P. Dempster, N. M. Laird, D. B. Rubin // Journal of the royal statistical society. – 1977. – Vol. 39, № 1. – pp. 1-39.

65. Dembo, A. Parameter estimation of partially observed continuous time stochastic processes via the EM algorithm / A. Dembo, O. Zeitouni // Stochastic Processes and their Applications. – 1986. – Vol. 23, № 1. – pp. 91-113.

66. Li, X. Training Hidden Markov Models with Multiple Observations – A Combinatorial Method / X. Li // IEEE Transactions on Pattern Analysis and Machine Intelligence. – 2000. – vol. PAMI 22, №4. – pp. 371-377.

67. Banfield, J. D. Model-Based Gaussian and Non-Gaussian Clustering / J. D. Banfield, A. E. Raftery // Biometrics. – 1993. – Vol. 49, №3. – pp. 803-821.

68. An Erratum for "A Tutorial on Hidden Markov Models" [Электронный ресурс] / A. Rahimi. – Режим доступа: http://alumni.media.mit.edu/~rahimi/rabiner/rabiner-errata/rabiner-errata.html#forward_scaling

69. Гультяева, Т. А. Вычисление первых производных от логарифма функции правдоподобия для скрытых марковских моделей / Т. А. Гультяева // Сборник Научных трудов НГТУ. – 2010. – №2. – С. 39-46.

70. Gultyaeva, T. A. Classification of observation sequences described by Hidden Markov Models / T. A. Gultyaeva, A. A. Popov, V. V. Kokoreva, V. E. Uvarov // Proceedings of the International Workshop Applied Methods of Statistical Analysis Nonparametric approach AMSA-2015. – 2015. – pp. 136-143.

71. Wang, H. Modeling Idea Generation Sequences Using Hidden Markov Models / H. Wang // Proceedings of the Annual Meeting of the Cognitive Science Society. – 2008. – №30. – pp. 107-112.

72. Уваров, В. Е. Анализ неполных последовательностей, описываемых скрытыми марковскими моделями / В. Е. Уваров, А. А. Попов, Т. А. Гультяева // Искусственный интеллект и принятие решений. – 2017. – №2. – С. 17-30.

73. Попов, А. А. Распознавание, декодирование и восстановление последовательностей с пропусками, описываемых скрытой марковской моделью с дискретным распределением наблюдений / А. А. Попов, Т. А. Гультяева, В. Е. Уваров // Научный вестник НГТУ. – 2017. – №1. – С. 99-119.

74. Uvarov, V. E. Imputation of Incomplete Motion Data Using Hidden Markov Models / V. E. Uvarov, A. A. Popov, T. A. Gultyaeva // Journal of Physics: Conference Series. – 2019. – 1210. – P. 012151.

75. Uvarov, V. E. Modeling multidimensional incomplete sequences using hidden Markov models / V. E. Uvarov, A. A. Popov, T. A. Gultyaeva // Proceedings of the International Workshop Applied Methods of Statistical Analysis Nonparametric approach AMSA-2017. – 2017. – pp. 343-349.

76. Quddus, M. A. Current map-matching algorithms for transport applications: State-of-the art and future research directions / M. A. Quddus, W. Y. Ochieng, R. B. Noland // Transportation Research Part C: Emerging Technologies. – 2007. – Vol. 15, №5. – pp. 312-328.

77. Algizawy, E. Real-Time Large-Scale Map Matching Using Mobile Phone Data / E. Algizawy, T. Ogawa, A. El-Mahdy // ACM Trans. Knowl. Discov. Data. – 2017. – Vol. 11, №4. – pp. 52:1-52:38.

78. OpenStreetMap contributors. Open Street Maps [Электронный ресурс] / OpenStreetMap contributors. – 2019. – Режим доступа: https:// www.openstreetmap.org

79. Rousseeuw, P. J. Alternatives to the median absolute deviation / P. J. Rousseeuw, C. Croux // Journal of the American Statistical Association. – 1993. – Vol. 88, №424. – pp. 1273–1283.

80. Gather, U. Robust estimation of scale of an exponential distribution / U. Gather, V. Schultze // Statistica Neerlandica. – 1999. – Vol. 53, №3. – pp. 327-341.

81. Eiter, T. Computing discrete Fréchet distance / T. Eiter, H. Mannila // Tech. Report CD-TR 94/64 at Christian Doppler Laboratory for Expert Systems. – 1994.

82. Уваров В. Е. Декодирование наиболее вероятного маршрута абонентов по транспортному графу на основе последовательности регистраций в мобильной сети / В. Е. Уваров, Д. В. Курганский, А. А. Попов, А. В. Климов, А. С. Мерзляков // T-Comm – Телекоммуникации и Транспорт. – 2019. – №7. – С. 32-39.

83. Popov, A. A. A survey of techniques for sequence recognition by using hidden Markov models when data loss occurs / A. A. Popov, V. E. Uvarov // Progress Through Innovations: тезисы городской научно-практической конференции аспирантов и магистрантов. – 2016. – pp. 32-33.

84. Uvarov, V. E. User Identification from Incomplete Motion Data Using Hidden Markov Models / V. E. Uvarov, A. A. Popov, T. A. Gultyaeva // 14th International Conference on Actual Problems of Electronic Instrument Engineering Proceedings (APEIE-2018). – 2018. – Vol. 1. – pp. 327-329.

85. Popov, A. A. Training Hidden Markov Models on Incomplete Sequences / A. A. Popov, T. A. Gultyaeva, V. E. Uvarov // 13th International Conference on Actual Problems of Electronic Instrument Engineering Proceedings (APEIE-2016). – 2016. – Vol. 1. – pp. 317-320.

86. Попов, А. А. Исследование подходов к обучению скрытых марковских моделей при наличии пропусков в последовательностях / А. А. Попов, Т. А. Гультяева, В. Е. Уваров // Материалы российской научно-технической

конференции «Обработка информации и математическое моделирование». – 2016. – С. 125-139.

87. Popov, A. A. A Comparison of Some Methods for Training Hidden Markov Models on Sequences with Missing Observations / A. A. Popov, T. A. Gultyaeva, V. E. Uvarov // Proceedings of 11th International Forum on Strategic Technology IFOST-2016. – 2016. – Vol. 1. – pp. 431-435.

88. Попов, А. А. Исследование Методов Обучения Скрытых Марковских Моделей при Наличии Пропусков в Последовательностях / А. А. Попов, Т. А. Гультяева, В. Е. Уваров // Труды XIII международной конференции Актуальные проблемы электронного приборостроения (АПЭП-2016). – 2016. – Т. 8. – С. 149-152.

89. Uvarov, V. E. A Survey of Techniques for Training Hidden Markov Models when Data Loss Occurs / V. E. Uvarov // Aspire to Science тезисы городской научнопрактической конференции студентов, магистрантов и аспирантов. – 2016. – pp. 127-128.

90. Уваров, В. Е. Обучение скрытых марковских моделей с непрерывной плотностью распределения наблюдений в условиях пропусков в последовательностях / В. Е. Уваров, А. А. Попов, Т. А. Гультяева // Сборник X Всероссийской научной конференции молодых ученых «НАУКА. ТЕХНОЛОГИИ. ИННОВАЦИИ». – 2016.

91. Уваров, В. Е. Обучение скрытых марковских моделей по неполным последовательностям / В. Е. Уваров, А. А. Попов, Т. А. Гультяева // Обработка информации и математическое моделирование ОИиММ-2017. – 2017.

92. Уваров В. Е. Распознавание неполных последовательностей, описываемых скрытыми марковскими моделями, в пространстве первых производных от логарифма функции правдоподобия / В. Е. Уваров // Вестник Томского Государственного Университета: управление, вычислительная техника и информатика. – 2018. – №42. – С. 79-88.

93. Гультяева, Т. А. Исследование возможностей применения алгоритма kближайших соседей и метода опорных векторов для классификации сигналов, порожденных скрытыми марковскими моделями / Т. А. Гультяева, Д. Ю. Коротенко // Материалы Всерос. науч. конф. молодых ученых Наука. Технологии. Инновации. – 2011. – Т. 1. – С. 242-246.

94. Гультяева, Т. А. Исследование возможностей применения алгоритма kближайших соседей и метода опорных векторов при классификации

последовательностей, порожденных скрытыми марковскими моделям / Т. А. Гультяева, Д. Ю. Коротенко // Сборник Научных трудов НГТУ. – 2011. – №3. – С. 45-55.

95. Cortes, C. Support-vector networks / C. Cortes, V. N. Vapnik // Machine Learning. – 1995. – №20. – pp. 273–297.

96. Rifkin, R. In Defense of One-Vs-All Classification / R. Rifkin, A. Klautau // Journal of Machine Learning Research. – 2004. – №5. – pp. 101-141.

97. Гультяева, Т. А. Классификация смоделированных скрытыми марковскими моделями последовательностей в многоклассовом случае / Т. А. Гультяева, А. А. Попов // Научный вестник НГТУ. – 2013. – №3. – С. 40-45.

98. Kohavi, R. A study of cross-validation and bootstrap for accuracy estimation and model selection / R. Kohavi // Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence. – 1995. – №2. – pp. 1137–1143.

99. Коротенко, Д. Ю. Построение гибридной модели для распознавания цифровых сигналов, основанной на комбинации скрытых марковских моделей и машин опорных векторов / Д. Ю. Коротенко, Т. А. Гультяева // Информатика и проблема телекоммуникаций: материалы российской науч.-технич. конф. – 2011. – Т. I. – С. 76-79.

100. Uvarov, V. E. Recognition of incomplete sequences using Fisher scores and hidden Markov models / V. E. Uvarov, A. A. Popov, T. A. Gultyaeva // Journal of Physics: Conference Series, XI International scientific and technical conference Applied Mechanics and Dynamics Systems. – 2018. – Vol. 944, №1. – P. 012121.

ПРИЛОЖЕНИЕ А АКТ О ВНЕДРЕНИИ РЕЗУЛЬТАТОВ ДИССЕРТАЦИОННОЙ РАБОТЫ

Общество с ограниченной ответственностью "Т2 Мобайл"
ИНН 7743895280 Адрес: 108811, Москва, Киевское ш., 22-й километр, д. 6, стр. 1 Телефон: (473) 258 00 65

АКТ

о внедрении результатов диссертационной работы Уварова Вадима Евгеньевича «Разработка и исследование методов распознавания последовательностей, описываемых скрытыми марковскими моделями, при неполных данных»

Настоящим актом подтверждается использование результатов диссертационного исследования Уварова Вадима Евгеньевича «Разработка и исследование методов распознавания последовательностей, описываемых скрытыми марковскими моделями, при неполных данных» компанией ООО “Т2 Мобайл” при разработке системы декодирования наиболее вероятного маршрута по транспортному графу на основе последовательности регистраций в мобильной сети. Результаты данного исследования, в частности разработанный впервые модифицированный алгоритм Витерби, позволили существенно повысить точность декодирования маршрутов по сравнению с ранее используемыми методами.

Директор по стратегическому планированию _____ / Скворцова С. А.

«___» _____ 2019 г.

ПРИЛОЖЕНИЕ Б СВИДЕТЕЛЬСТВО О ГОСУДАРСТВЕННОЙ РЕГИСТРАЦИИ ПРОГРАММЫ ДЛЯ ЭВМ

РОССИЙСКАЯ ФЕДЕРАЦИЯ

RU 2017615226

ФЕДЕРАЛЬНАЯ СЛУЖБА
ПО ИНТЕЛЛЕКТУАЛЬНОЙ СОБСТВЕННОСТИ

ГОСУДАРСТВЕННАЯ РЕГИСТРАЦИЯ ПРОГРАММЫ ДЛЯ ЭВМ

Номер регистрации (свидетельства):
2017615226

Дата регистрации: **05.05.2017**

Номер и дата поступления заявки:
2017612064 14.03.2017

Дата публикации: **05.05.2017**

Контактные реквизиты:
+7-913-792-43-58,
uvarov.vadim42@gmail.com

Автор:
Уваров Вадим Евгеньевич (RU)

Правообладатель:
Уваров Вадим Евгеньевич (RU)

Название программы для ЭВМ:
Анализатор неполных последовательностей, описываемых скрытыми марковскими моделями

Реферат:
Программа предназначена для анализа неполных последовательностей с помощью алгоритмов, основанных на использовании теории скрытых марковских моделей (СММ). Неполные последовательности могут состоять как из наблюдений, принадлежащих дискретному множеству символов, так и из многомерных векторов действительных чисел, при этом значение некоторых наблюдений может быть не определено. В программе реализуются алгоритмы обучения СММ по неполным последовательностям, а также алгоритмы распознавания, восстановления и декодирования неполных последовательностей с помощью СММ.

Тип реализующей ЭВМ:	IBM PC-совмест. ПК
Язык программирования:	Python 3.5
Вид и версия операционной системы:	Windows XP/7/8/8.1/10, Unix
Объем программы для ЭВМ:	84 Кб