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

АВТОРЕФЕРАТ

диссертации на соискание ученой степени

кандидата технических наук

Новосибирск – 2019

Работа выполнена в Федеральном государственном бюджетном образовательном учреждении высшего образования «Новосибирский государственный технический университет», г. Новосибирск.

Научный руководитель: **Попов Александр Александрович,**
доктор технических наук, профессор

Официальные оппоненты: **Рябко Борис Яковлевич,**
доктор технических наук, профессор,
Федеральное государственное бюджетное учреждение науки Институт вычислительных технологий Сибирского отделения Российской академии наук, г. Новосибирск, главный научный сотрудник;

Каргаполова Нина Александровна,
кандидат физико-математических наук,
Федеральное государственное бюджетное учреждение науки Институт вычислительной математики и математической геофизики Сибирского отделения Российской академии наук, г. Новосибирск, научный сотрудник

Ведущая организация: Федеральное государственное бюджетное учреждение науки Институт проблем управления им. В. А. Трапезникова Российской академии наук, г. Москва

Защита состоится «27» февраля 2020 г. в 16^{00} часов в конференц-зале на заседании диссертационного совета Д 212.173.06 при Федеральном государственном бюджетном образовательном учреждении высшего образования «Новосибирский государственный технический университет» по адресу: 630073, г. Новосибирск, пр. К. Маркса, 20.

С диссертацией можно ознакомиться в библиотеке Новосибирского государственного технического университета и на сайте http://nstu.ru.

Автореферат разослан «___» января 2020 г.

Ученый секретарь
диссертационного совета  Андрей Владимирович Фаддеенков

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

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 (L. E. Baum, T. Petrie, J. A. Eagon, G. R. Sell). Традиционно СММ применялись для распознавания речи, например, в работах M. Gales, S. Young, S. E. Levinson, L. R. Rabiner, M. M. Sondhi, J. K. Baker, L. R. Bahl, D. P. Munteanu, S. A. Toma, M. J. Gales. Начиная с 1980-х годов СММ стали применять в биоинформатике (см. работы E. Birney, M. J. Bishop, E. A. Thompson, J. Söding, L. Käll, A. Krogh, E. L. Sonnhammer), например, при анализе цепочек ДНК (см. работы S. A. Malekpour, H. Pezeshk, M. Sadeghi). Также СММ успешно применялись для моделирования экономических процессов (см. работы R. Bhar, S. Hamori, C. Erlwein, R. Mamon, M. Davison, R. E. McCulloch, R. S. Tsay, A. Rossia, G. M. Gallob) и в задачах компьютерного зрения (см. работы H. Bunke, T. Caelli, A. V. Nefian, M. H. Hayes, F. Niu, M. Abdel-Mottaleb, L. Zhang, Y. Chen, G. Fang, X. Chen, W. Gao). Наибольшей популярностью СММ стали пользоваться после 1990-х годов, и данная тенденция сохранилась вплоть до настоящего времени, что можно подтвердить частотой упоминания термина “hidden Markov model” в публикациях. Одним из недавних способов применения СММ для моделирования, являются задачи распознавания двигательной активности человека. Этот класс задач включает в себя как распознавание совершаемого движения (см. работы K. Altun, B. Barshan, O. Tunçel, B. Barshan, M. C. Yüksek, K. Altun, B. Barshan), так и идентификацию субъекта, совершающего движение (см. работы P. Casale, O. Pujol,

P. Radeva, C. Nickel, C. Busch, H. Brandt). Также СММ хорошо зарекомендовали себя при решении задач декодирования оптимального маршрута по последовательности геоданных (см. работы P. Newson, J. Krumm, H. Koller, P. Widhalm, M. Dragaschnig, A. Graser, C. Y. Goh, J. Dauwels, N. Mitrovic, M. T. Asif, A. Oran, P. Jaillet, R. Mohamed, H. Aly, M. Youssef, G. Wang, R. Zimmermann).

Проблема использования СММ для анализа неполных последовательностей частично освещается в статье авторов M. Cooke, P. Green, L. Josifovski и A. Vizinh, 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 работе авторов D. Lee, D. Kulic, Y. Nakamura. 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” 3-6 октября 2016 года в г. Новосибирск; на международной конференции “Прикладные методы статистического анализа: непараметрические подходы в кибернетике и системном анализе” в г. Красноярск 17-22 сентября 2017 года; на XI Международной IEEE научно-технической конференции “Динамика систем, механизмов и машин” в г. Омск 14-16 ноября 2017 года; на XII Международной IEEE научно-технической конференции “Динамика систем, механизмов и машин” в г. Омск 13-15 ноября 2018 года; на международной конференции “Прикладные методы статистического анализа: непараметрический подход” (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) 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» 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 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

Во введении обоснована актуальность и отражена степень разработанности темы исследования, сформулированы цели и задачи диссертационной работы, выделены объект и предмет исследования, изложена идея работы и её научная новизна, приведены теоретическая и практическая значимость исследования, перечислены основные положения, выносимые на защиту, указана реализация результатов работы, описан личный вклад автора, а также приведены структура и объём научно-квалификационной работы.

В первой главе исследовано современное состояние проблемы анализа последовательностей, описываемых скрытыми марковскими моделями. Выявлены пробелы теории в области анализа последовательностей, содержащих пропуски.

Изучены основные понятия теории «скрытых марковских моделей», на которых базируются теоретические методы, разработанные автором. Согласно теории приняты обозначения: N число скрытых состояний модели, S – скрытое состояние модели, T - длина последовательности, t - момент времени, q_t – скрытое состояние в момент времени t, $\boldsymbol{o}_t$ – наблюдение в момент времени t, O – последовательность наблюдений, M – количество компонент нормальных распределений в смеси. В сущности, СММ можно представить в виде набора определяющих её параметров $\lambda = \{\Pi, A, B\}$, где параметр Π соответствует вектору вероятностного распределения начального скрытого состояния $\Pi = \left\{\pi_i = p\left(q_1 = s_i\right),\ i = \overline{1,N}\right\}$, параметр A соответствует матрице 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 параметр B соответствует множеству условных плотностей распределения наблюдений в скрытых состояниях (вероятности эмиссии): $B = \left\{b_i\left(\boldsymbol{o}\right) = p\left(\boldsymbol{o} \mid q = s_i\right), i = \overline{1,N}\right\}$. В случае СММ с непрерывной плотностью распределения, вероятности $b_i\left(\boldsymbol{o}\right)$ описываются смесями нормальных распределений: $b_i(\boldsymbol{o}) = \sum_{m=1}^{M} \tau_{im} g(\boldsymbol{o}; \mu_{im}, \Sigma_{im}),\ \ i = \overline{1,N}, \boldsymbol{o} \in R^Z$.

Рассмотрены современные решения задачи декодирования последовательностей, описываемых скрытыми марковскими моделями, заключающиеся в установлении наиболее вероятной последовательности скрытых состояний $\hat{Q} = \{\hat{q}_1, \ldots, \hat{q}_T\}$, в котором находился случайный процесс, сгенерировавший последовательность. Установлено, что наиболее применим алгоритм Витерби, решающий данную задачу.

Изучены методы распознавания последовательностей, описываемых СММ. Выбран оптимальный алгоритм расчёта значения функции правдоподобия $p\left(O \mid \lambda\right)$ под названием forward-backward (прямой-обратный), с помощью которого можно проводить распознавание по критерию максимума функции правдоподобия (далее – МФП).

Рассмотрены методы обучения скрытых марковских моделей, а именно, эффективный итеративный алгоритм Баума-Велша для обучения СММ, который является частным случаем алгоритма EM (expectation-maximization; ожидание-максимизация). Приведены рекомендации по выбору начальных приближений параметров СММ.

Описан приём масштабирования вероятностей в формулах анализа последовательностей, описываемых СММ, позволяющий производить анализ длинных последовательностей без возникновения проблемы переполнения вещественных переменных вычислительной машины.

Установлено, что первые производные от логарифма функции правдоподобия того, что описываемый СММ процесс сгенерировал последовательность, можно использовать в качестве векторов признаков, описывающих последовательности. С помощью этих признаков возможно проводить их классификацию, применяя классические алгоритмы машинного обучения, например, метод опорных векторов.

Применительно к исследуемым алгоритмам изучено понятие «неполная последовательность наблюдений». Неполная или «дефектная» последовательность – это такая последовательность, в которой значение некоторых наблюдений не определено (т. е. имеются пропуски). При этом наличие пропусков определяется некоторыми внешними факторами: то есть изучаемый процесс порождает 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 $\varnothing$. Приведены методики моделирования целых и неполных последовательностей, описываемых скрытыми марковскими моделями.

Первая глава завершается постановкой задач исследования.

Во второй главе приведен разработанный автором метод восстановления и декодирования неполных последовательностей, описываемых СММ.

Автором предложена новая формула вероятности эмиссии для случая пропущенного наблюдения $b_i(\varnothing),\ i=\overline{1,N}$ с помощью приёма маргинализации, в использовании которого заключается основная идея диссертационной работы. Маргинализация – это приём использования маргинального распределения, т. е. распределения некоторых случайных величин без указания на значения других случайных величин. В сущности, данный приём заключается в интегрировании вероятности по всем возможным значениям неизвестной величины. После применения приёма маргинализации получаются следующие формулы вероятности эмиссии:

$$b_i(\varnothing)=\sum_{v\in V} b_i(v)=1,\ i=\overline{1,N}$$ для дискретного распределения наблюдений; (1)

$$b_i(\varnothing)=\int b_i(\boldsymbol{x})\,d\boldsymbol{x}=1,\ i=\overline{1,N}$$ для непрерывного распределения наблюдений. (2)

При этом формула плотности нормального распределения, входящего в смесь, для наблюдения-пропуска примет вид:

$$g(\varnothing,\mu_{im},\Sigma_{im})=\int g(\boldsymbol{x},\mu_{im},\Sigma_{im})\,d\boldsymbol{x}=1,\ \begin{matrix} i=\overline{1,N} \\ m=\overline{1,M} \end{matrix}. \quad (3)$$

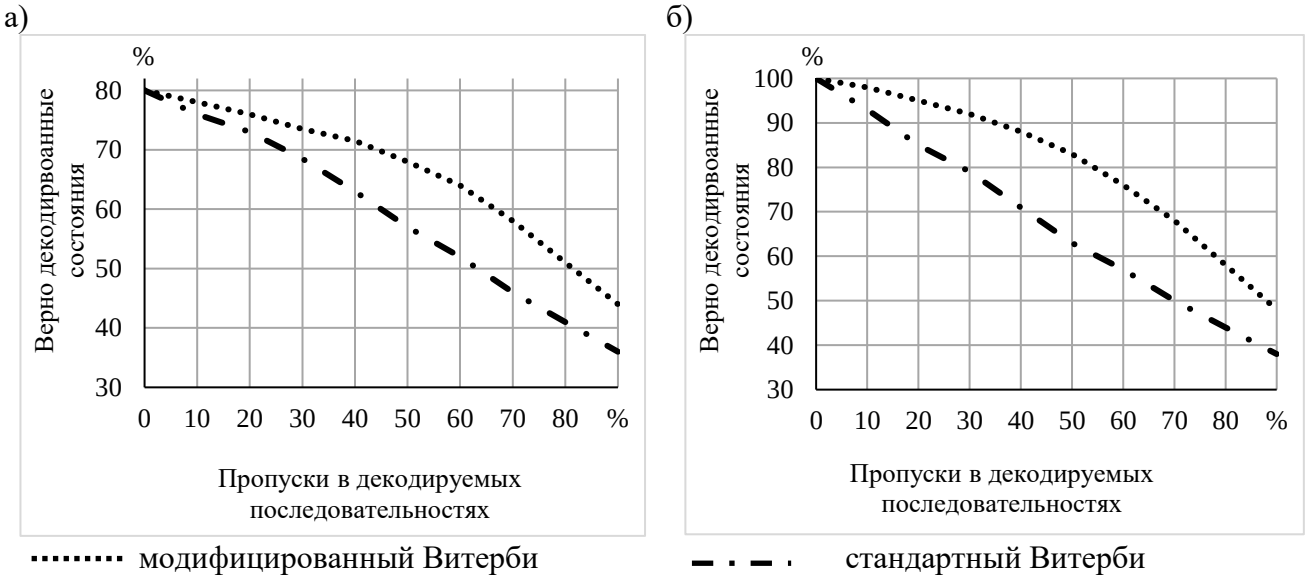
С использованием полученных выше формул (1)-(3) решена задача **декодирования** неполной последовательности, описываемой СММ, на основе алгоритма Витерби, модифицированного для случая появления пропусков в последовательностях.

Для **восстановления** неполных последовательностей в новом методе вначале проводится декодирование неполной последовательности модифицированным алгоритмом Витерби, а затем на основе декодированных скрытых состояний, соответствующих пропущенным наблюдениям, производится генерация замещающих наблюдений.

Для сравнения рассмотрен стандартный метод **восстановления** неполных последовательностей, основанный на замещении пропусков некоторой статистикой от значений соседних наблюдений. В случае СММ с дискретным распределением пропуски замещаются модой соседних наблюдений, а в случае СММ с непрерывным распределением – средним арифметическим соседних наблюдений.

Стандартный же метод **декодирования** заключается в восстановлении последовательности стандартным методом с последующим применением обычного алгоритма Витерби.

Проведена оценка эффективности разработанного метода путем **декодирования** последовательностей, сгенерированных с помощью заданной СММ, в которых случайным образом были введены пропуски. На рисунке 1 даны графики зависимостей количества верно декодированных состояний от пропусков для модифицированного и стандартного алгоритмов Витерби.

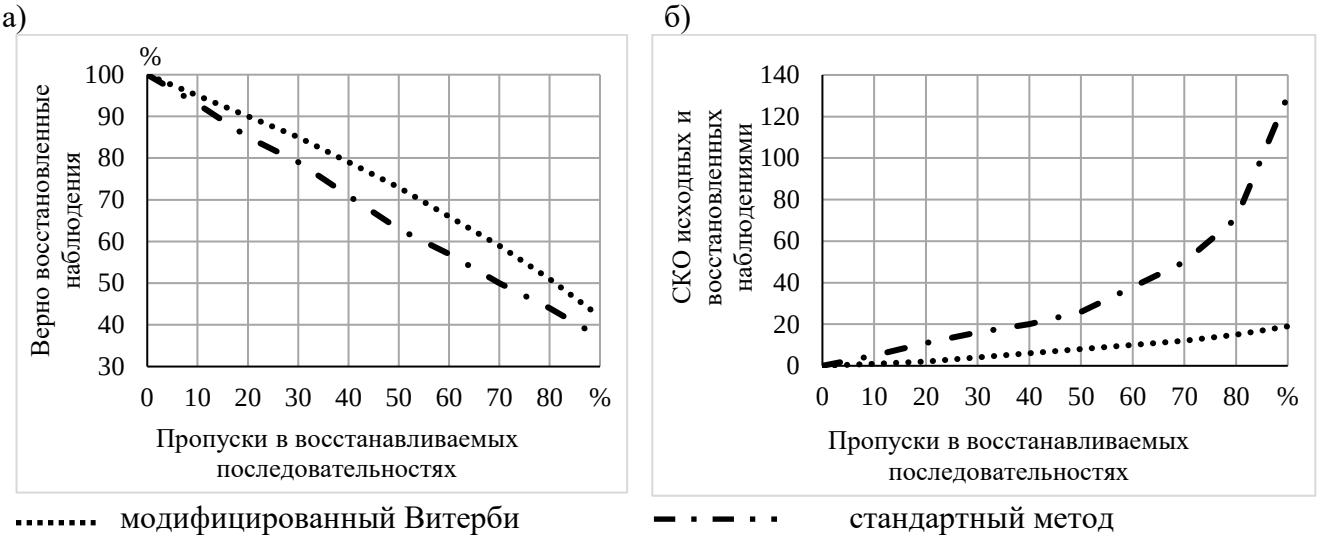


а) дискретные наблюдения, б) вектора вещественных чисел

Рисунок 1 – Эффективность модифицированного алгоритма Витерби для декодирования неполных последовательностей

В обоих случаях на рисунке 2, для дискретного (а) и непрерывного (б) распределений наблюдений из графиков видно преимущество применения в новом методе модифицированного алгоритма Витерби над стандартным методом (до 1.4 раз).

Проведена оценка эффективности разработанного метода, основанного на модифицированном алгоритме Витерби, для **восстановления** неполных последовательностей, сгенерированных с помощью заданной СММ, в которых случайным образом были введены пропуски. Рисунок 2 содержит сравнение стандартного метода и метода на основе модифицированного алгоритма Витерби.



а) дискретные наблюдения, б) вектора вещественных чисел

Рисунок 2 – Эффективность модифицированного алгоритма Витерби для восстановления неполных последовательностей

В обоих случаях на рисунке 2, для дискретного (а) и непрерывного (б) распределений наблюдений из графиков видно преимущество применения в новом методе модифицированного алгоритма Витерби над стандартным методом (до 1.4 раз).

На основании теоретических исследований автором предложена методика восстановления неполных данных двигательной активности человека с применением модифицированного алгоритма Витерби. На первом этапе проводится обучение СММ по целым последовательностям, характеризующим двигательную активность. Затем с помощью СММ и модифицированного алгоритма Витерби восстанавливаются новые неполные последовательности.

Для проверки разработанной методики использовался набор данных “User Identification From Walking Activity” свободно доступный в интернете. В наборе данных содержится информация, генерируемая смартфоном на базе операционной системы Android, расположенным в нагрудном кармане. Снимались показатели акселерометра телефона, в то время как каждый из участников эксперимента шёл по определённому заранее маршруту. Всего в эксперименте участвовало 22 человека. Результаты проведённого эксперимента по оценке эффективности решения данной задачи новым и стандартным методами представлены на рисунке 3. Для обучения было использовано 75% целых последовательностей из выборки, а восстановление проводилось на 25% оставшимся последовательностям, в которых предварительно были сгенерированы пропуски. Из графика на рисунке 3 видно значительное преимущество (до 5 раз) предложенной методики над стандартным методом, особенно при большом проценте пропусков в данных.

Автором также предложена методика декодирования наиболее вероятного пути движения абонента по транспортному графу на основе последовательности регистраций в мобильной сети. Задача заключается в превращении информации о регистрациях абонента на базовых станциях мобильной связи в траекторию его движения по улицам города. Схематичный пример решения подобной задачи представлен на рисунке 4. Здесь 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.

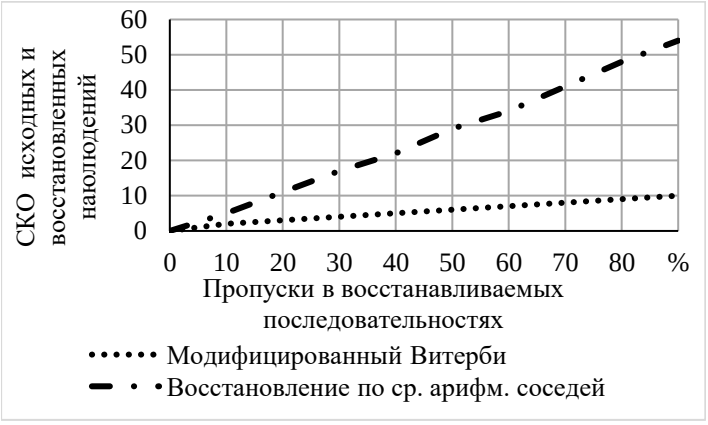


Рисунок 3 – Эффективность нового метода для восстановления неполных данных двигательной активности человека

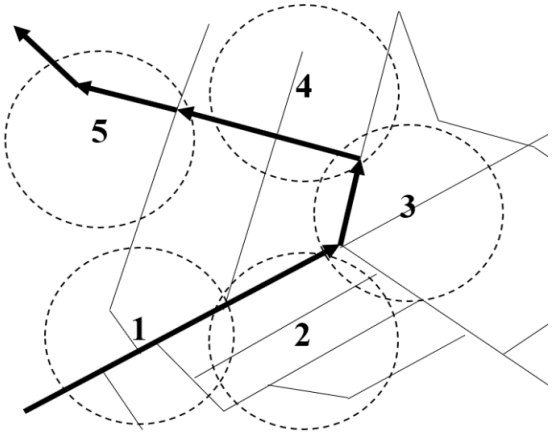


Рисунок 4 – Пример фрагмента транспортного графа с наложенными на него покрытиями секторов, на которых последовательно регистрировался абонент

Суть методики заключается в том, что вершины транспортного графа моделируются как скрытые состояния СММ, а регистрации абонента в мобильной сети – как наблюдения. Вероятности эмиссий рассчитываются обратно пропорционально расстоянию от регистрации до базовой станции мобильной сети, а вероятности переходов – обратно пропорционально длине пути по графу между вершинами. При такой постановке задачи для нахождения оптимального пути движения абонента по графу достаточно решить задачу декодирования последовательности.

В этом случае неполнота данных проявляется в том, что между некоторыми наборами вершин, соответствующим паре последовательных регистраций невозможно найти ни одного пути по графу (из-за ограничений на максимальную длину пути, связанную с производительностью). Предложенная методика решает эту проблему с помощью того же приёма маргинализации и модифицированного алгоритма Витерби: вероятностям переходов между каждой парой вершин присваивается значение 1, в результате чего они не влияют на итоговый выбор оптимального маршрута.

Для оценки качества предложенной методики был проведён вычислительный эксперимент на 20000 анонимизированных суточных треков абонентов, 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 декодируется отдельно с помощью стандартного алгоритма Витерби, а во втором случае используется модифицированный алгоритм Витерби для устранения таких разрывов в маршрутах. Данные Таблица 1 содержат измеренное количество точек трека-образца, для которых нашлась соответствующая точка на восстановленном треке. 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.

Таблица 1 – Эффективность модифицированной версии 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

Во втором эксперименте проведено сравнение алгоритма декодирования оптимального маршрута, использующего скрытые марковские модели (модифицированный Витерби) 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.

Как видно из таблицы 2 алгоритм, использующий вероятностный подход на основе СММ, позволяет увеличить точность более чем в 2.5 раза по сравнению с простым алгоритмом, соединяющим координаты БС, и не учитывающим транспортный граф.

Таблица 2 – Эффективность модифицированной версии алгоритма Витерби для построения наиболее вероятного пути движения абонента по транспортному графу

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

Предложенная методика успешно внедрена в платформу анализа геоданных компании

ООО “Т2 Мобайл” (мобильный оператор Tele2), что подтверждается соответствующим актом о внедрении.

В третьей главе приведен разработанный автором метод распознавания неполных последовательностей, описываемых СММ.

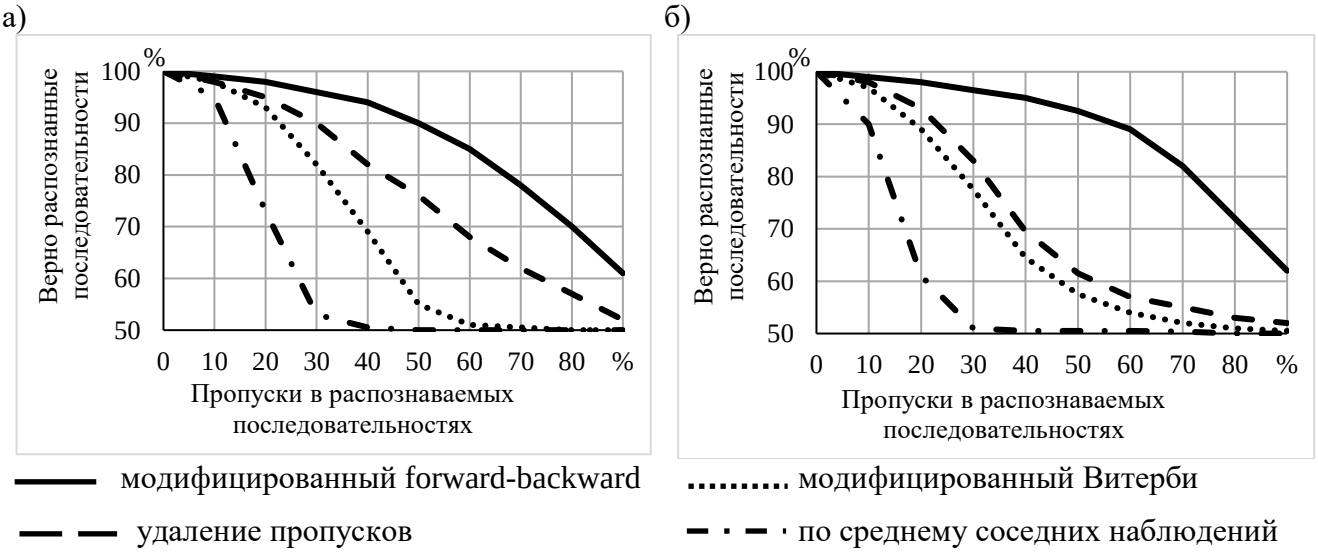
На основе формулы вычисления вероятности эмиссии для пропущенного наблюдения (1)-(3) модифицирован алгоритм forward-backward для случая появления пропусков в последовательностях. Распознавание неполных последовательностей далее проводится с помощью критерия МФП.

Для сравнения использовали методы: удаление пропусков, восстановление по модифицированному алгоритму Витерби, а также по моде или среднему арифметическому соседних наблюдений.

Метод с модифицированным алгоритмом Витерби состоит в том, что пропуски в последовательности вначале восстанавливают с помощью метода на основе модифицированного алгоритма Витерби, а затем распознают с помощью стандартного алгоритма forward-backward по критерию МФП. Значение функции правдоподобия вычисляют с помощью той СММ, по которой проводилось восстановление.

Метод удаления пропусков заключается в том, что вначале из неполной последовательности исключаются пропуски, а оставшиеся подпоследовательности «склеиваются» между собой и далее распознаются стандартного алгоритма forward-backward по критерию МФП.

Результат исследования эффективности разработанного метода распознавания неполных последовательностей на основе модифицированного алгоритма forward-backward представлен на рисунке 5.



а) дискретные наблюдения, б) вектора вещественных чисел

Рисунок 5 – Эффективность модифицированного алгоритма forward-backward для распознавания неполных последовательностей

В обоих случаях на рисунке 5, для дискретного (а) и непрерывного (б) распределений наблюдений видно преимущество разработанного алгоритма распознавания, основанного на модифицированном алгоритме forward-backward, над другими результатами (до 1.6 раз).

На основании теоретических исследований автором предложена практическая методика идентификации личности по неполным данным двигательной активности при полной обучающей выборке. Вначале обучают несколько СММ по **целым** последовательностям, характеризующим двигательную активность, а затем с помощью полученных СММ, каждая из которых соответствует манере движения индивидуальной личности, проводят распознавание новых неполных последовательностей. Для проверки разработанной методики использовали набор данных двигательной активности человека, описанный выше. На рисунке 6 показана зависимость количества правильно распознанных неполных последовательностей от пропусков в этих последовательностях. Эффективность методики сравнивали по аналогичному набору методов (рисунок 5). Из диаграммы на рисунке 6 видно значительное преимущество разработанного метода распознавания, основанного на модифицированном алгоритме forward-backward, особенно при увеличении пропусков.

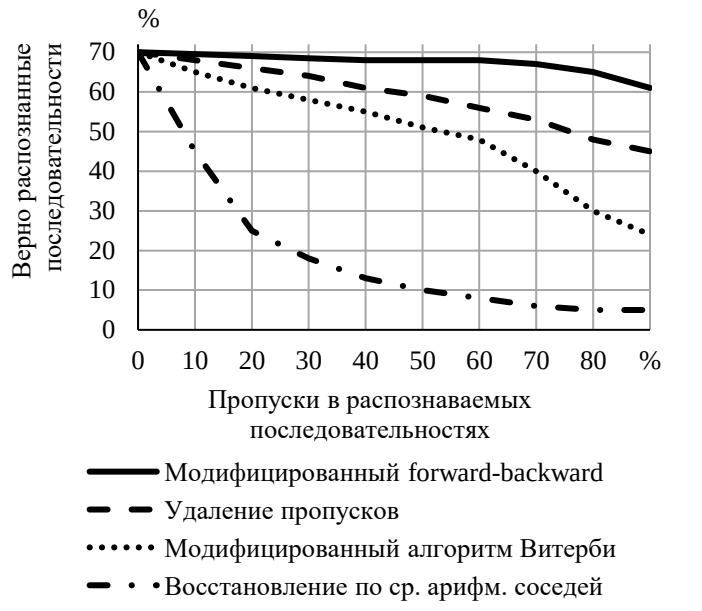


Рисунок 6 – Эффективность модифицированного алгоритма forward-backward для идентификации личности при неполных данных двигательной активности

В четвёртой главе приведен разработанный автором метод обучения СММ по неполным последовательностям. На основе формул вычисления вероятности эмиссии для пропущенного наблюдения (1)-(3) модифицирован алгоритм Баума-Велша, что делает его устойчивым к пропускам.

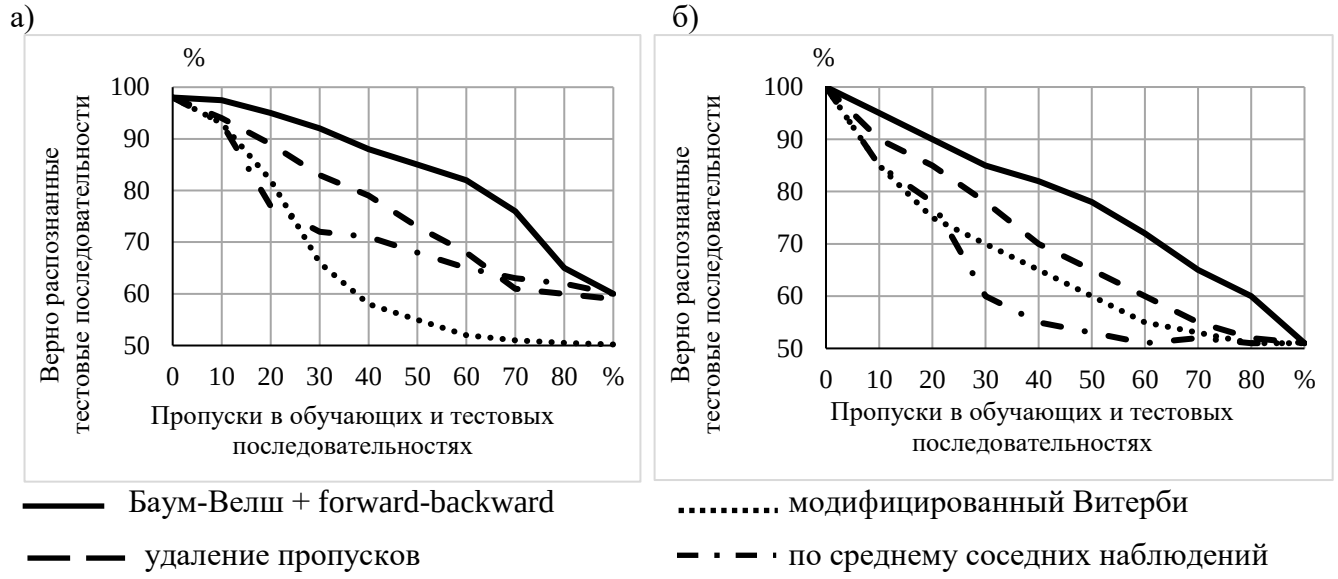
В эксперименте оценивалась эффективность использования СММ, обученных по неполным последовательностям, в качестве классификаторов таких последовательностей. Сравнивались следующие методы:

а) обучение с помощью модифицированного алгоритма Баума-Велша (новый метод) и распознавание с помощью модифицированного алгоритма forward-backward по критерию МФП;

б) обучения и распознавания стандартными алгоритмами путём предварительного исключения пропусков из последовательностей;

в) обучения и распознавания стандартными алгоритмами путём предварительного восстановления неполных последовательностей с помощью модифицированного алгоритма Витерби;

г) обучение и распознавание стандартными алгоритмами путём предварительного восстановления последовательностей с пропусками по моде или среднему соседних наблюдений. Результаты эксперимента представлены на рисунке 7.



а) дискретные наблюдения, б) вектора вещественных чисел

Рисунок 7 – Эффективность модифицированного алгоритма Баума-Велша для построения классификаторов неполных последовательностей

В обоих случаях (рисунок 7) видно преимущество разработанного метода обучения СММ по неполным последовательностям, основанного на модифицированном алгоритме Баума-Велша над остальными методами (до 1.2 раз).

На основе теоретических исследований автором разработана методика идентификации личности по неполным данным двигательной активности при неполной обучающей выборке. Решение задачи заключается в том, чтобы сначала обучить несколько СММ по **неполным** последовательностям, характеризующим двигательную активность, а затем с помощью полученных СММ, каждая из которых соответствует манере движения индивидуальной личности, проводить распознавание новых неполных последовательностей.

На рисунке 8 показаны результаты проведённого эксперимента по оценке эффективности решения данной задачи новым и стандартными методами. Эффективность методики сравнивали по аналогичному набору методов (рисунок 7). Из графика на рисунке 8 видно преимущество разработанного алгоритма обучения СММ по неполным последовательностям, основанного на модифицированном алгоритме Баума-Велша, над остальными алгоритмами.

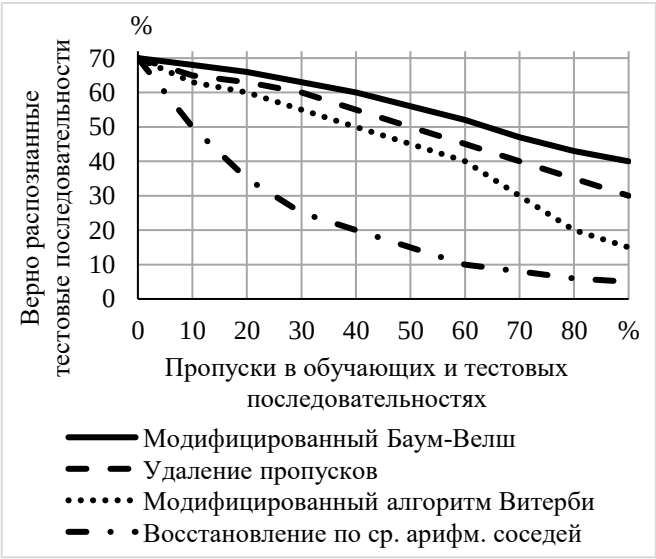


Рисунок 8 – Эффективность методики идентификации личности по неполным данным двигательной активности при неполной обучающей выборке

В пятой главе описан новый метод распознавания неполных последовательностей, порождённых близкими по параметрам СММ. Он основан на модифицированном алгоритме

вычисления первых производных от функции правдоподобия того, что случайный процесс, описываемый СММ, сгенерировал неполную последовательность. Для модификации алгоритма также использовались формулы вычисления вероятности эмиссии для пропущенного наблюдения (1)-(3).

Решение задачи распознавания неполных последовательностей, описываемых близкими по параметрам СММ, базируется на использовании метода опорных векторов для классификации неполных последовательностей на основании соответствующих им векторов признаков, образованных из производных от логарифма функции правдоподобия, вычисленных с помощью модифицированного алгоритма.

Исследована эффективность разработанного метода распознавания по первым производным от логарифма функции правдоподобия в сравнении с методом распознавания, основанным на модифицированном алгоритме forward-backward. Результаты эксперимента представлены на рисунке 9.

Оценки СММ были получены модифицированным алгоритмом Витерби по неполным обучающим последовательностям. Далее с помощью вычисленных производных от неполных обучающих последовательностей был обучен классификатор метода опорных векторов. Распознавание проводилось по производным от тестовых последовательностей с использованием полученного классификатора.

Разработанный метод распознавания сравнивался с методом на основе модифицированного алгоритма forward-backward с классификатором по критерию МФП (с использованием тех же оценок СММ).

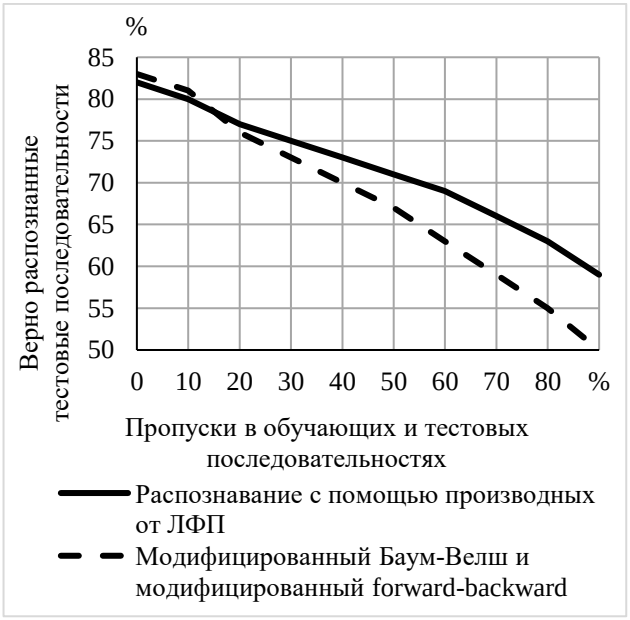


Рисунок 9 – Оценка эффективности метода распознавания неполных последовательностей в пространстве первых производных от логарифма функции правдоподобия

Как видно из рисунка 9, метод распознавания, основанный на производных, начинает превосходить метод, основанный на алгоритме forward-backward, начиная примерно с 20% пропусков в обучающих и тестовых последовательностях. При этом достигается увеличение точности в 1.2 раза при 90% пропусков в последовательностях.

На основании вышеизложенных теоретических исследований автором разработана практическая методика идентификации личности по неполным данным двигательной активности с использованием производных от логарифма функции правдоподобия по параметрам СММ. Поскольку двигательная активность многих людей достаточно схожа, была выдвинута гипотеза, что алгоритм классификации, основанный на производных, потенциально может увеличить точность идентификации. Суть методики заключается в следующем. Вначале обуча-

ются несколько СММ по неполным последовательностям, характеризующим двигательную активность. Затем с помощью полученных СММ, каждая из которых соответствует манере движения индивидуальной личности, проводится распознавание новых неполных последовательностей по алгоритму, основанному на производных.

Для проверки разработанной методики использовался набор данных двигательной активности, описанный ранее. Эффективность методики сравнивали по аналогичному набору методов (рисунок 9). Результаты проведённого эксперимента представлены на рисунке 10. Как видно из рисунка 10, метод распознавания, основанный на производных, превосходит метод, основанный на модифицированном алгоритме forward-backward. Причём при увеличении процента пропусков в обучающих и тестовых последовательностях, различие в точности также увеличивается. Таким образом, метод, основанный на производных, рекомендуется применять при близких по параметрам конкурирующих СММ.

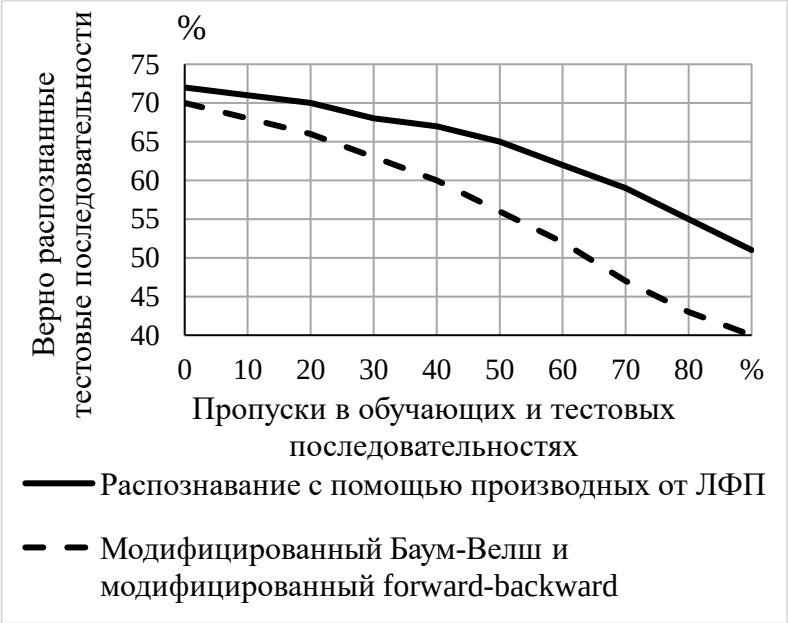


Рисунок 10 – Эффективность методики распознавания личности по неполным данным с помощью производных от логарифма функции правдоподобия

ОСНОВНЫЕ РЕЗУЛЬТАТЫ ИССЛЕДОВАНИЯ

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 последовательно-

стей 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 точность распознавания последовательностей с помощью обученных моделей до 1.2 раз при дискретном и непрерывном распределении наблюдений, чем при использовании других 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 точность распознавания подобных последовательностей до 1.2 раза 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 СММ до 5 раз;

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

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

ПУБЛИКАЦИИ ПО ТЕМЕ ДИССЕРТАЦИИ

Публикации в изданиях из «Перечня российских рецензируемых научных журналов, в которых должны быть опубликованы основные научные результаты диссертаций на соискание ученых степеней доктора и кандидата наук» ВАК

1. 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.

2. 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.

3. 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.
4. 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. – Т.13, №7. – С. 32 39.

Публикации в изданиях, входящих в базы данных Scopus и Web Of Science

5. Popov, A. A. Training Hidden Markov Models on Incomplete Sequences / A. A. Popov, T. A. Gultyaeva, V. E. Uvarov // Proceedings of Conference on Actual Problems of Electronic Instrument Engineering. – 2016. – Vol. 1. – pp. 317-320.
6. Uvarov, V. E. Modeling multidimensional incomplete sequences using hidden Markov models / V. E. Uvarov, A. A. Popov, T. A. Gultyaeva // International Workshop Applied Methods of Statistical Analysis Nonparametric approach. – 2017. – pp. 343-349.
7. 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.
8. Uvarov, V. E. User Identification from Incomplete Motion Data Using Hidden Markov Models / V. E. Uvarov, A. A. Popov, T. A. Gultyaeva // Conference on Actual Problems of Electronic Instrument Engineering Proceedings. – 2018. – Vol. 1. – pp. 327-329.
9. 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.

Публикации в сборниках научных работ и материалах конференций РИНЦ

10. 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.
11. 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.
12. 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 // International Forum on Strategic Technology. – 2016. – Vol. 1. – pp. 431-435.
13. 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.
14. Uvarov, V. E. A Survey of Techniques for Training Hidden Markov Models when Data Loss Occurs / V. E. Uvarov // Aspire to Science тезисы научно практической конференции студентов, магистрантов и аспирантов. – 2016. – pp. 127-128.

15. 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. – С. 194 196.
16. Уваров, В. Е. Обучение скрытых марковских моделей по неполным последовательностям / В. Е. Уваров, А. А. Попов, Т. А. Гультяева // Материалы российской научно-технической конференции «Обработка информации и математическое моделирование» ОИиММ-2017. – 2017. – С. 169-177.

Свидетельство о государственной регистрации для ЭВМ
17. Уваров, В. Е. Свидетельство №2017615226 о государственной регистрации для ЭВМ «Анализатор неполных последовательностей, описываемых скрытыми марковскими моделями» / В. Е. Уваров // Федеральный институт промышленной собственности. – 2017.

Отпечатано в типографии Новосибирского
государственного технического университета
630073, г. Новосибирск, пр. К. Маркса, 20
Тел./факс (383) 346-08-57
Формат 60×84×1/16. Объем 1.5 п.л. Тираж 100 экз.
Заказ №150. Подписано в печать 12.12.2019 г.