На правах рукописи

Нурутдинова Алсу Рафаиловна

ИДЕНТИФИКАЦИЯ И КЛАССИФИКАЦИЯ АВТОМАТНЫХ МАРКОВСКИХ МОДЕЛЕЙ МЕТОДАМИ МНОГОПАРАМЕТРИЧЕСКОГО АНАЛИЗА

01.01.09 – дискретная математика и математическая кибернетика

АВТОРЕФЕРАТ

диссертации на соискание ученой степени
кандидата физико-математических наук

Казань – 2018

Работа выполнена на кафедре «Компьютерные системы» ФГБОУ ВО «Казанский национальный исследовательский технический университет им. А. Н. Туполева – КАИ»

Научный руководитель: **Шалагин Сергей Викторович**,
доктор технических наук, доцент,
ФГБОУ ВО «Казанский национальный исследовательский технический университет им. А. Н. Туполева – КАИ»

Официальные оппоненты: **Кирпичников Александр Петрович,**
доктор физико-математических наук, профессор,
ФГБОУ ВО «Казанский национальный исследовательский технологический университет»

Крашенинников Виктор Ростиславович, доктор технических наук, профессор,
ФГБОУ ВО «Ульяновский государственный технический университет»

Ведущая организация: ФГБОУ ВПО «Национальный исследовательский Нижегородский государственный университет им. Н.И. Лобачевского»

Защита состоится «20» декабря 2018 г. в 14.30 на заседании диссертационного совета Д. 212.081.35 при федеральном государственном автономном образовательном учреждении высшего образования «Казанский (Приволжский) федеральный университет (ФГАОУ ВО КФУ), по адресу: 420008, г. Казань, ул. Кремлевская, д.35, Институт математики и механики им. Н. И. Лобачевского, аудитория 1011.

С диссертацией можно ознакомиться в Научной библиотеке им. Н. И. Лобачевского ФГАОУ ВО «Казанский (Приволжский) федеральный университет» по адресу: 420008, г. Казань, ул. Кремлевская, д. 35. Электронная версия диссертации размещена на официальном сайте Казанского (Приволжского) федерального университета http://kpfu.ru.

Автореферат разослан «___» _____ 2018 года.

Ученый секретарь

диссертационного совета Д 212.081.35

кандидат физико-математических наук, доцент А.И. Еникеев

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы исследования. Цепи Маркова (ЦМ) и их функции используются для моделирования процессов, событий, явлений, имеющих вероятностную (стохастическую) природу, динамических сложных систем, объектов из области экономики, социологии, медицины и для анализа поведения технических устройств. Популярность марковских моделей объясняется тем, что они дают возможность получить адекватное описание широкого класса процессов, событий и явлений, а также относительной простотой: для ЦМ будущее состояние при известном текущем состоянии не зависит от состояний в прошлом. Вопросы анализа энтропийных и асимптотических свойств дискретных марковских процессов исследовались в работах отечественных (Альпин Ю.А., Бухараев Р.Г., Гиоргадзе А.Х., Глова В.И., Захаров В.М., Лоренц А.А., Королюк В.С., Романовский В.И., Федотов Н.Г. и др.) и зарубежных ученых (Джадж Д., Зельнер А., Кемени Дж., Ли И., Рабинер Л., Снелл Дж. и др.)

Известным направлением математической кибернетики является задача распознавания вероятностных автоматных моделей в заданном семействе по наблюдаемым состояниям моделируемой системы. Актуальность ее связана с тем, что получаемые результаты распознавания необходимы для принятия решений в системах управления. В частности, в теоретических и прикладных исследованиях важную роль играют задачи статистического распознавания, связанные с классификацией и идентификацией реализаций экспериментов на выходе автоматных моделей, разрешимостью распознавания определенных классов автоматных моделей, анализом эффективности методик и алгоритмов, оценок их сложности, достаточных для распознавания.

В работах Барашко А.С., Павлива А.Н. решены задачи статистического распознавания детерминированных автоматов по их выходным и по вход-выходным последовательностям при условии, что на их входах действуют генераторы стохастических последовательностей. Однако вопрос статистического распознавания (идентификации) системы «генератор стохастических последовательностей – детерминированный автомат» к априори заданным подклассам исследован не достаточно.

Известный метод решения задачи идентификации ЦМ (И.Ли, Д.Джадж, А.Зельнер и др.) характеризуется высокой погрешностью биграмм, триграмм и т.п., построенных на основе ЦМ, относительно эргодических стохастических матриц (ЭСМ), задающих генератор дискретных стохастических марковских процессов.

Актуальность решения задач классификации марковских моделей в работах Захарова В.М., Нурмеева Н.Н., Салимова Ф.И., Соколова С.Ю., Шалагина С.В. (2001-2003 гг.) обусловлена проблемой представимости генераторов конечных однородных ЦМ на основе ограниченных ресурсов ПЛИС. Аппаратные затраты для хранения множества файлов конфигурации ПЛИС возрастают при увеличении числа данных генераторов ЦМ. Ими предложены методы многопараметрической классификации автоматных марковских моделей (АММ) с целью уменьшения объема исходных данных для моделирования цепей Маркова с заданными свойствами.

Вероятностные автоматы используются при решении задач классификации и распознавания (идентификации) речи, как письменной, так и устной (Соловьев В.Д., Сулейманов Д.Ш. и др.), но вопросы применения для данной цели методов многопараметрического анализа изучены не достаточно.

Предложенные модели и методы классификации и идентификации ЦМ на основе функционалов, определяемых на основе биграмм, характеризуются тем, что они вычисляются с определенной погрешностью относительно стохастических матриц, задающих ЦМ, что снижает точность анализа ЦМ при использовании данных функционалов. Особенно – для ЦМ длины порядка $10^2 – 10^3$. Поэтому требуется разработка новых подходов к распознаванию, эффективных при анализе выходных последовательностей, реализаций ЦМ, с учетом длины фрагмента распознавания, размерности стохастических матриц и точности представления элементов ЦМ. Требуют исследования вопросы анализа состава кластеров при решении задач классификации ЦМ.

Актуальность темы работы обусловлена потребностью создания математических моделей распознавания АММ, повышением эффективности методов в соответствии с выделением определенных подклассов исследуемых объектов, созданием новых алгоритмов распознавания, повышением точности результатов статистического анализа.

В задачах анализа важным является выбор информативных признаков классификации. Анализ множества классификационных признаков и выявление наиболее информативных позволяет улучшить временные характеристики алгоритмов классификации, а также качество получаемых решений. Другой важной задачей является исследование вероятности идентификации ЦМ в зависимости от длины их реализаций, генерируемых на основе АММ. Для повышения качества идентификации АММ требуются эффективные методы, ориентированные на определенную структуру ЭСМ, задающих АММ.

Полученные в работе результаты позволяют с большей вычислительной эффективностью классифицировать и идентифицировать (распознавать) АММ по выходной последовательности, используя интегрированные признаки на основе фрагмента последовательности, но в то же время и с учетом структуры ЭСМ. Оценена степень достоверности в зависимости от длины фрагмента и, в некоторых случаях, от размерности ЭСМ. Исследованы две независимые модели распознавания АММ: первая основана на идентификации эргодических ЦМ и вторая − циклических ЦМ.

Объект исследования: дискретные стохастические процессы, порождаемые на основе автоматных марковских моделей. Предмет исследования: алгоритмы, применяемые для классификации и идентификации автоматных марковских моделей при использовании многопараметрического анализа с применением заданного множества признаков.

Цель работы: идентификация и классификация различных подклассов АММ на основе генерируемых дискретных ЦМ, при использовании разработанных моделей, методик и алгоритмов.

Эффективность идентификации и классификации АММ определяется снижением длины ЦМ, требуемых для решения задачи идентификации и классификации АММ, определенных на основе заданных подклассов ЭСМ, с определенной доверительной вероятностью, а также уменьшением вычислительной сложности алгоритмов распознавания и снижением погрешности вычисления признаков относительно ЭСМ.

В соответствии с поставленной целью были решены следующие задачи:

1. Разработка математической модели и методики идентификации циклической ЦМ на основе последовательности ее состояний конечной длины. Исследовано применение модификации алгоритма «прямого-обратного хода» к решению задачи идентификации АММ, определенных на основе циклической ЭСМ.
2. Модификация модели и алгоритма «прямого-обратного хода» для идентификации конечных простых однородных ЦМ, сгенерированных на основе ЭСМ. Решена задача идентификации конечных простых однородных ЦМ, часть элементов которой скрыта от наблюдения.
3. Разработка алгоритмов: а) многопараметрической классификации АММ, задаваемых на основе ЭСМ, принадлежащих к определенным подклассам; б) идентификации априори задаваемых подклассов АММ, определяемых при использовании подклассов ЭСМ, на основе последовательностей состояний ЦМ конечной длины.
4. Описание подхода для анализа состава кластеров, выделяемых путем многопараметрической классификации множества АММ, определяемых различными группами признаков.
5. Создание комплекса программ, реализующих указанные алгоритмы анализа и идентификации автоматных марковских моделей.

Методы исследований. Для решения поставленных задач использованы методы и понятия теории вероятностей, теории случайных процессов, методы статистической обработки данных, теории множеств, теории автоматов, дискретной математики.

В работе получены следующие результаты, характеризующиеся научной новизной:

1. Предложен подход решения задачи идентификации конечных простых однородных ЦМ, сгенерированных на основе ЭСМ определенных подклассов. В том числе – ЭСМ из класса циклических. Предложена модификация модели и алгоритма Л. Рабинера, позволяющая идентифицировать принадлежность ЦМ к заданному подклассу, определяемому структурой ЭСМ. Причем часть элементов ЦМ может быть скрыта от наблюдения.
2. Разработаны алгоритмы: а) идентификации принадлежности АММ к подклассу, задаваемому структурой задающей ее ЭСМ, на основе порождаемых ими ЦМ определенной длины с возможностью вычисления вероятности корректной идентификации; б) многопараметрической классификации множества АММ, определяемых ЭСМ, принадлежащих различным подклассам, на основе генерируемых ими ЦМ заданной длины; в) анализа состава кластеров, выделяемых в результате кластеризации АММ и/или порождаемых ими цепей Маркова на основе заданного множества признаков.
3. Создан комплекс программ, основу которого составляют связанные по входным данным алгоритмы и программы, служащие для реализации разработанных численных методов идентификации, многопараметрической классификации и анализа АММ.

На защиту выносятся следующие основные результаты, полученные в ходе диссертационной работы:

1. Математическая модель и методики идентификации длины цикла ЦМ с получением доверительной вероятности корректной идентификации.
2. Модификация модели Л. Рабинера для вычисления вероятности идентификации АММ на основе порождаемой ею последовательности состояний ЦМ конечной длины, в том числе последовательности со скрытыми состояниями.
3. Алгоритмы: а) многопараметрической классификации АММ, определенных на основе задающих их ЭСМ, принадлежащих к заданным подклассам, имеющим разную степень различия структур; б) идентификации с заданной доверительной вероятностью АММ, определяемых ЭСМ, принадлежащими к одному из априори заданных подклассов, на основе последовательности состояний ЦМ конечной длины, порождаемой указанной АММ.
4. Численный метод анализа состава кластеров, позволяющий определить их общее количество кластеров во множестве объектов классификации − АММ, на основе предложенных критериев, характеризующих дисперсию элементов внутри каждого из кластеров.
5. Комплекс программ, реализующих предложенные численные методы анализа: классификации и идентификации АММ.

Теоретическая и практическая значимость состоит в разработке новых моделей для идентификации и классификации АММ, а также в развитии численных методов анализа указанных моделей. Результаты, изложенные в диссертации, могут найти применение при решении задач распознавания различных процессов и систем, определяемых и/или описываемых на основе АММ. Полученные результаты позволяют классифицировать и идентифицировать только по выходной последовательности, используя признаки на основе не одного элемента n-граммы $n = 2, 3, \ldots$, а последовательности фрагментов ЦМ. При этом множество признаков рассчитывается с учетом особенностей ЭСМ, то есть, исследована сходимость по вероятности элементов признаков, вычисленных на основе ЦМ конечной длины N, к соответствующим элементам признаков, вычисленных на основе ЭСМ. Оценена степень достоверности идентификации в зависимости от длины ЦМ и, в случае циклической ЦМ, от размерности ЭСМ.

Апробация работы. Основные результаты работы были доложены и обсуждались на конференциях и семинарах *международного уровня*: «Будущее технической науки» (Нижний Новгород, 2007), «Туполевские чтения» (Казань, 2007, 2008, 2009, 2011), «Инфокоммуникационные технологии глобального информационного общества» (Казань, 2009), межд. школа-семинар «Синтез и сложность управляющих систем» им. академ. О.Б. Лупанова (Пенза, 2009), «Проблемы техники и технологий телекоммуникаций» (Казань, 2008), «Инфокоммуникационные технологии глобального информационного общества» (Казань, 2009), «Проблемы теоретической ки-

бернетики» (Нижний Новгород, 2011), «Актуальные проблемы и перспективы развития гражданской авиации России» (Иркутск, 2016), Новые технологии, материалы и оборудование российской авиакосмической отрасли (Казань, 2016 г.) , «Проблемы анализа и моделирования региональных социально-экономических процессов» (Казань, 2017); *всероссийского* уровня: «Наука технологии Инновации» (Нижний Новгород, 2007), «Информационные технологии в системе социально – экономической безопасности России и её регионов» (Казань, 2009, 2010), «Информационные технологии-2010» (Йошкар-Ола, 2010), «Динамика нелинейных дискретных электротехнических и электронных схем» (Чебоксары, 2017); *регионального* уровня: «Наука и профессиональное образование» (Нижнекамск, 2007).

Результаты исследований использовались при разработке программ для ЭВМ [23-26].

Публикации. По теме диссертации опубликовано 27 работ: 7 статей в ведущих рецензируемых научных изданиях, 4 св-ва о регистрации программ для ЭВМ и 16 работ в сборниках трудов и материалов конференций и семинаров международного (10), всероссийского (5) и регионального (1) уровней.

Сведения о личном вкладе автора. Предложена модель классификации АММ [2] и идентификации АММ [1]. Разработаны численные методы классификации [2,8-14], идентификации АММ [1, 3, 5-7, 15,16,18] и анализа структуры кластеров [17]. Разработаны алгоритмы [4, 20-23] и комплекс программ многопараметрического анализа АММ [24-27].

Структура и объем работы: введение, четыре главы, заключение и список используемых источников, включающий 85 наименований. Объем работы – 140 стр. Работа включает 31 рисунок и 15 таблиц.

ОСНОВНОЕ СОДЕРЖАНИЕ ДИССЕРТАЦИИ

Во введении обоснована актуальность исследований, проводимых в рамках диссертационной работы, сформулированы цели и задачи исследования.

Первая глава содержит вспомогательные сведения из областей распознавания образов, случайных процессов и автоматов, необходимые для дальнейшего изложения в следующих главах. Рассмотрена математическая модель, базирующаяся на определениях цепи Маркова и вероятностного автомата. Приведен обзор существующих подходов к анализу вероятностных автоматных моделей.

Определение 1 (Кемени Дж., Снелл Дж., 1970). Простая однородная цепь Маркова задана в виде

$$\left(S, P_s, \pi_0 \right), \tag{1}$$

где $S = \{s_i\}$, $i = \overline{0, n-1}$ – множество состояний ЦМ, P_s – эргодическая стохастическая матрица вида $P_s = \left(p_{ij}\right)$ размерности *n*x*n*, $i, j = \overline{0, n-1}$, а π_0 – *n*-мерный вектор начального распределения вероятности появления каждого состояния ЦМ.

Определение 2 (Поспелов Д.А., 1970). Автономным вероятностным автоматом будем называть систему

$$\left(S, \varphi(s'/s) \right), \tag{2}$$

где $s, s' \in S$, $\varphi(s', s)$ – функция переходов, заданная стохастической матрицей P_s.

В качестве АММ(*P*), порождающей ЦМ, будем рассматривать автомат вида (2), заданный на основе (1) по алгоритму разложения ЭСМ *P* (Поспелов Д.А, 1970). На АММ(*P*) накладывается ограничение: значения положительных элементов *P* кратны заданной величине *D*.

Приведены описания методов многопараметрической классификации данных, применяемых для решения задач статистического анализа АММ. Выполнен обзор работ по анализу АММ и ЦМ, который позволил сформулировать актуальность задачи, подходы к разработке новых алгоритмов распознавания. Дается постановка задач диссертации.

Входные данные. 1. Пусть $AMM(P)$ – АММ вида (2), заданная на основе ЭСМ P. **2.** Задано множество подклассов ЭСМ $P - Q$, $AMM(P \in Q)$, которые выделены в зависимости от характеристик ЭСМ *P*. Число подклассов задано априори. На всем множестве реализаций ЦМ существует разбиение на подмножества (подклассы ЭСМ *P*). **3.** Для $AMM(P \in Q)$ получено множество *S* реализаций ЦМ длины *N* вида $\hat{S}(N) = s_1, s_2 ... s_N$, где s_t – состояние АММ вида (2) в момент времени *t*, $t = \overline{1, N}$. **4.** Задано множество классификационных эмпирических признаков $\tilde{h} = \{x_i\}, i = 1,..., L$, вычисленных на основе реализаций ЦМ и множество теоретических признаков $h = \{y_j\}, j = 1,..., M$, вычисленных на основе ЭСМ *P*. Совокупность признаков для $\hat{S}(N)$ определяется множеством $\hat{h}(\hat{S}) = (x_1(\hat{S}), x_2(\hat{S}),..., x_L(\hat{S}))$, $h(\hat{S}) = (y(\hat{S}), y_2(\hat{S}),..., y_M(\hat{S}))$.

Если АММ определена как «черный ящик», когда неизвестна функция перехода $\varphi(s', s)$ АММ, но известны последовательности $\hat{S}(N)$ то можно рассматривать задачи классификации и идентификации АММ по реализациям ЦМ.

В случае задачи идентификации требуется сформировать эффективное множество признаков подклассов Q_i АММ P $\tilde{h} = \{x_i\}, i = 1,..., L$ и построить алгоритм (решающую функцию, методику) а: $S \to Q$, способный отнести последовательность $\hat{S}(N) \in S$ длины *N* к подклассу Q_i с использованием множества признаков $\tilde{h}$. Для каждого объекта $\hat{S}(N)$ необходимо определить значения α, вероятности того, что $\hat{S}(N)$ сгенерирована на основе $AMM(P)$, где ЭСМ *P* принадлежит заданному подклассу Q_i.

В случае задачи классификации требуется: 1) определить критерии $K(h(\hat{S}))$, позволяющие отличать $\hat{S}(N)$, генерируемых на основе АММ, принадлежащих разным подклассам; 2) определить минимальную длину последовательностей $\hat{S}(N)$, достаточную для классификации с заданной доверительной вероятностью; 3) построить алгоритм $b: S \to Q$, способный классифицировать (различать) последовательность $\hat{S}(N) \in S$ в соответствии с заданным критерием, включающий в себя композиции методов, средств классификации и методику их использования; 4) построить разбиение множества *S* на классы Q_i. Соответствие критерия определяется следующим образом: $\forall i, j, i \neq j \Leftrightarrow K(h(\hat{S} \in Q_i)) \neq K(h(\hat{S} \in Q_j))$.

Задача классификации АММ разбивает множества $\hat{S}(N)$ на непересекающиеся классы, позволяя таким образом получить характеристики выделенных кластеров и их типичных представителей, выявить информативные признаки, определить алгоритмы, способные различать объекты исследования.

Вторая глава посвящена описанию предложенной математической модели и методик идентификации и классификации автоматных марковских моделей на основе ЭСМ.

В п. 2.1 предложена модель многопараметрической классификации с использованием частотных признаков для АММ, определенных на основе ЭСМ, принадлежащих к определенным подклассам и задающих указанные АММ [1]. Задано множество реализаций ЦМ $\hat{S}(N) = s_1, s_2 ... s_N$, полученных на основе АММ(*P*), где *P* принадлежит одному из подклассов – $Q_i = \{T_н, T_в, Б_л, Б_п\}$: квазитреугольные (*T*), нижние и верхние, и блочно-сообщающиеся (*Б*), правые и левые. Выбор данных подклассов ЭСМ обоснован тем, что каждый из них имеет различную степень взаимного сходства-различия по структуре. Определено множество классифицирующих признаков $\hat{h}$, позволяющее с заданной вероятностью разделить АММ, заданные на базе ЭСМ из различных подклассов [9, 10, 11, 21]. Предложен набор эмпирических признаков, кото-

рые характеризуют частоту перехода ЦМ $\hat{S}(N) = s_1, s_2 ... s_N$ из состояния s_i в состояние s_j ($s_i, s_j \in S$, $i, j = \overline{0, n-1}$) для $i - j = k$, рассчитывается по формуле вида

$$\tilde{f}_k = \frac{1}{N-1}\sum_{z=1}^{N-1}\hat{y}_z,\ \hat{y}_z = \begin{cases} 1: & y_z = k \\ 0: & y_z \neq k \end{cases},\quad k = \overline{1-n, n-1}\quad . \tag{3}$$

Вероятности перехода ЦМ f_k ($k = \overline{1-n, n-1}$) из состояния s_i в состояние s_j для случая, когда $i - j = k$, являющиеся теоретическими оценками для величин $\tilde{f}_k$, определяются на базе ЭСМ *P*, определяющих АММ(*P*) и рассчитываются по формуле

$$f_k = \frac{1}{n}\sum_{i=1}^{n-|k|} m(i,k), \tag{4}$$

где $m(i,k) = p_{i(i-k)}$ для $k = \overline{1-n, 0}$ и $m(i,k) = p_{(i+k)i}$ для $k = \overline{1, n-1}$.

Предлагаемая модель определяет новый подход к решению задачи многопараметрической классификации АММ на основе генерируемых ими последовательностей состояний при использовании метода дискриминантного анализа (ДА), состоящего из следующих этапов [8]: 1) вычисление границ области определения каждого признака вида (3), определяющих критерии $K(h(\hat{S}))$; 2) выявление признаков, значимых при выделении того или иного подкласса ЭСМ *P*, задающих АММ(*P*) на основе $K(h(\hat{S}))$; 3) вычисление для заданной $\hat{S}(N)$ элементов рабочего множества (РМ) $\tilde{f}_k$ вида (4) с априори указанной точностью, при этом определяется минимальная длина *N* последовательности, снимаемой с выхода АММ, которая требуется для достижения заданной точности ε_k; 4) разделение множества реализаций ЦМ, снимаемых с выхода АММ(*P*), на группы Q_i, определяемые подклассами ЭСМ P, $P \in \{T_н, T_в, Б_л, Б_п\}$, при использовании ДА. Минимальная длина последовательности зависит от отклонения теоретических признаков от эмпирических при гипотезе нормального распределения[1]:

$$\varepsilon_k = \left| f_k - \tilde{f}_k \right| = t_\beta \sqrt{\frac{\tilde{f}_k\left(1-\tilde{f}_k\right)}{N-1}},\ k = \overline{1-n, n-1}, \tag{5}$$

если $\tilde{f}_k \in \left[\frac{5}{N-1}, \frac{N-6}{N-1}\right]$; в противном случае при биномиальном законе $\varepsilon_k = \max\left(\left|\tilde{f}_k - p_{1k}\right|, \left|p_{2k} - \tilde{f}_k\right|\right)$, где p_{1k} и p_{2k} - верхняя и нижняя границы доверительного интервала, рассчитанные при заданном значении доверительной вероятности β: $\sum_{m=l}^{N-1} C_{N-1}^m p_{1k}^m (1-p_{1k})^{N-m-1} = \frac{\beta}{2}$, $\sum_{m=0}^{l} C_{N-1}^m p_{2k}^m (1-p_{2k})^{N-m-1} = \frac{\beta}{2}$, где l - количество переходов ЦМ из состояния s_i в состояние s_j для $i - j = k$, $k = \overline{1-n, n-1}$ [2]. Методика отбора наиболее значимых для дискриминации признаков, основана на следующих утверждениях [6]:

Утверждение 1 [4]. Признаки f_k вида (4), объединенные во множество R_Q, являются значимыми для дискриминации АММ($P \in Q$) и включаются в рабочее множество, если диапазоны их значений не пересекаются: $f_k(Q) = 0$, а $\min f_k(\neg Q) > 0$, $k \in J_Q$. $f_k(Q)$ - значение *k*-го признака, вычисленного для подкласса ЭСМ *Q*, где *Q* принадлежит одному и только одному из заданных подклассов $P \in \{T_н, T_в, Б_л, Б_п\}$. J_Q - множество индексов признаков $\tilde{f}_k$, для которых $f_k(Q) = 0$.

Утверждение 2 [4]. Объединение множеств признаков f_k вида (4) $R_{Q^{(1)}}$ и $R_{Q^{(2)}}$, значимых для дискриминации АММ($P_1 \in Q^{(1)}$) и АММ($P_2 \in Q^{(2)}$), $Q^{(1)} \cap Q^{(2)} = \varnothing$, есть искомое РМ.

Утверждения имеют следующее практическое значение: множество признаков классификации возможно редуцировать путем отбора из него только значимых признаков.

Замечание 1 [4]. РМ для дискриминации АММ(P) на основе генерируемых ими ЦМ, $P \in \{T_н, T_в, Б_л, Б_п\}$ - $R = R_{Tн} \cup R_{Tв} \cup R_{Бп}$. Максимально допустимое отклонение ε, вычисленное согласно (5), равно $(n-1)Dn^{-1}$.

Минимальное количество испытаний, необходимых для выделения подклассов реализаций ЦМ, порождаемых АММ(P), определяется на основе (5) для каждого из признаков РМ согласно неравенству [2]

$$N = \max_{f_k \in R}\left(\tilde{f}_k\left(1-\tilde{f}_k\right)t_\alpha^2\varepsilon_k^{-2}+1\right) \underset{(\tilde{f}=0{,}5)}{\leq} \frac{n^2 t_\alpha^2}{4(n-1)^2 D^2}+1\ , \tag{6}$$

где ε_k определяется как максимально допустимое отклонение $\tilde{f}_k \in R$ от f_k, определенное при решении второй задачи - $(n-1)Dn^{-1}$.

Замечание 2 [4]. В случае, когда условие $\tilde{f}_k \in \left[\frac{5}{N-1}, \frac{N-6}{N-1}\right]$ не выполняется для признаков R с множеством индексов J, минимальное число испытаний для $\tilde{f}_k$, $k \in J$, определяется путем решения системы уравнений относительно N_k:

$$\begin{cases}(n-1)Dn^{-1} \geq \tilde{f}_k - p_{1k} \\ (n-1)Dn^{-1} \geq p_{2k} - \tilde{f}_k\end{cases} \text{ или } \begin{cases}p_{1k} \geq \tilde{f}_k - (n-1)Dn^{-1} \\ p_{2k} \leq (n-1)Dn^{-1} + \tilde{f}_k\end{cases}, \tag{7}$$

где p_{1k} и p_{2k} определяются при $\alpha = 0{,}95$ согласно выражению:

$$\sum_{m=l}^{N_k-1} C_{N_k-1}^m p_{1k}^m \left(1-p_{1k}\right)^{N_k-m-1} = 0{,}475,\ \sum_{m=0}^{l} C_{N_k-1}^m p_{2k}^m \left(1-p_{2k}\right)^{N_k-m-1} = 0{,}475,$$

l - количество переходов ЦМ из состояния s_i в s_j для случая, когда $i - j = k$, $k = \overline{1-n, n-1}$.

Экспериментальные данные подтверждают гипотезу относительно способности признаков, характеризующих частоту перехода ЦМ из состояния s_i в состояние s_j при $i - j = k$ ($s_i, s_j \in S$, $i, j = \overline{0, n-1}$), классифицировать последовательности, генерируемые АММ(P). На основе признаков, входящих в РМ, проведено разделение полученных 400 последовательностей, сгенерированных на основе АММ(P), элементы которой варьируются с точностью представления $D = 5 \cdot 10^{-2}$, на основе метода ДА. Определена длина последовательностей, достаточная для дискриминации порождающих их АММ(P) на априори заданные группы с заданной доверительной вероятностью [1]. Минимальное количество испытаний N, необходимых для выделения подклассов реализаций ЦМ при n=5 не превышает 564. Увеличение размерности n ЭСМ способствует снижению данной оценки. Доля верно дискриминированных последовательностей при использовании заданного РМ, составляет 100% [1].

В п. 2.2 предложен численный метод многопараметрической идентификации АММ, которые описаны ЭСМ, принадлежащими к априори заданным подклассам [2]. Выделены признаки, позволяющие идентифицировать подкласс ЭСМ P, которая определяет АММ(P), по реализациям ЦМ [12-15], и сведены в следующие группы, на основе их способности выделять определенные свойства ЦМ заданной длины N при некоторых ограничениях на P: 1) функционалы на основе m-грамм, для m = *2, 3* (V и W соответственно); 2) частотные признаки (F); 3) элементы биграммы (M).

Признаки из группы *V* определяются системой [4]

$$v_g = \begin{cases} 1: P \in X(g) \\ 0: \textit{иначе} \end{cases},\ g = \overline{1,4} \tag{8}$$

где $X(1) = Т_и$, $X(2) = Т_в$, $X(3) = Б_п$ и $X(4) = Б_л$. Принадлежность *P* к подклассу $X(g)$ для v_g, $g = \overline{1,4}$, определена априори. Признаки v_g, определенные согласно (8), являются теоретическими оценками величин $\tilde{v}_g$ вычисление которых производится путем подсчета частот по биграмме $\tilde{P} = (\tilde{p}_{ij})_{n \times n}$, построенной на основе ЦМ определенной длины [2]:

$$\tilde{v}_g = \begin{cases} 1: \forall \tilde{p}_{ij} : \left(\left(\tilde{p}_{ij} \geq 0 \right) \wedge \left(p_{ij} > 0 \right) \right) \wedge \left(\left(\tilde{p}_{ij} = 0 \right) \wedge \left(p_{ij} = 0 \right) \right) \wedge \left(P \in X(g) \right) \\ 0: \textit{иначе} \end{cases}, \tag{9}$$

где $i, j = \overline{0, n-1}$, $g = \overline{1,4}$, значение $X(g)$ определено по аналогии с (8), $\tilde{p}_{ij} = \frac{1}{N} \sum_{z=1}^{N-1} (s(z) = i) \,\&\, (s(z+1) = j)$, $s(z)$ - состояние ЦМ в дискретный момент времени z.

Необходимая длина *N* для идентификации вычисляется согласно формуле:

$$N' = \frac{\tilde{h}\left(1 - \tilde{h}\right)}{\left| h - \tilde{h} \right|^2} t_\alpha^2 \leq \max_{\tilde{h}} \left(\frac{\tilde{h}\left(1 - \tilde{h}\right)}{\varepsilon^2} t_\alpha^2 \right) = \frac{1}{4\varepsilon^2} t_\beta^2, \tag{10}$$

где ε - отклонение h и $\tilde{h}$, t_β - величина, определяющая для нормального закона число среднеквадратических отклонений, которое нужно отложить относительно центра рассеивания, чтобы вероятность попадания в полученный интервал была равна α.

$$\varepsilon_F = \min_{l,\, z=1,4,\, l \neq z} \left| f_k - \tilde{f}_k \right| = n^{-1}(n-1)D,\ k = \overline{1-n, n-1}. \tag{11}$$

Для группы *M* минимальное отклонение каждого из признаков – $p_{ij}^{(1)}$ от $p_{ij}^{(2)}$, $i, j = \overline{1,n}$, равно *D*. Для признаков группы *V*

$$\varepsilon_V = \min_{l,\, z=1,4,\, l \neq z} \left(\frac{1}{n} \sum_{i=1}^{n} r_i \right), \text{ где } r_i = \begin{cases} 1: & \textit{если } \forall p_{ij}^{(1)} = 0,\ p_{ij}^{(2)} > 0 \\ q_{lz}(i) \cdot D: & \textit{иначе} \end{cases}, \tag{12}$$

где $p_{ij}^{(1)}$ и $p_{ij}^{(2)}$ - элементы ЭСМ P_1 и P_2, $q_{lz}(i)$ - количество элементов в *i*-й строке ЭСМ P_1, для которых $p_{ij}^{(1)} = 0$, а $p_{ij}^{(2)} > 0$. Для признаков группы *W*

$$\varepsilon_W = \min_{l,\, z=1,4,\, l \neq z} \left(\frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} r'_{ij} \right), \text{ где } r'_{ij} = \begin{cases} 1: & \textit{если } \forall p_{ijr}^{(1)} = 0,\ p_{ijr}^{(2)} > 0 \\ q'_{lz}(i,j) \cdot D^2: & \textit{иначе} \end{cases}, \tag{13}$$

где $p_{ijr}^{(1)}$ и $p_{ijr}^{(2)}$ - элементы трехмерных массивов P'_1 и P'_2, образованных на основе ЭСМ P_1 и P_2, $q'_{lz}(i,j)$ - количество элементов $p_{ijr}^{(1)} \in P'_1$, для которых $p_{ijr}^{(1)} = 0$, а $p_{ijr}^{(2)} > 0$.

Теорема 1 [2]. Верхняя оценка длины подпоследовательности N', на основе которой вычисляются значения $\tilde{h}$, достаточной для идентификации АММ(P_1) и АММ(P_2), где P_1 и P_2 принадлежат различным подклассам ЭСМ с доверительной вероятностью α, определяется согласно формуле (10).

Теорема обосновывает численный метод решения задачи идентификации подкласса АММ(*P*), $P \in \{Т, Б\}$ и включает четыре этапа [2]: 1) определение *h*, $h \in \{v, w, f, m\}$ - рабочего множества; 2) на основе ЭСМ $P \in \{Т, Б\}$ вычисление ε_h; 3) определение длины ЦМ – N_h; 4) проведение идентификации подкласса АММ(*P*) для признаков группы *h* по ЦМ длины N_h на основе метода ДА в интегрированной системе Statistica 8.0 [9].

Предложенный численный метод позволяет для АММ (*P*), $P \in \{T, Б\}$, определить признаки для групп *M, F* и *V*, которые являются максимально эффективными при решении задачи идентификации заданной АММ(*P*), где *P* имеет заданную размерность и принадлежит заданному подклассу из множества $\{T_н, T_в, Б_л, Б_п\}$, и минимальную длину ЦМ, порождаемой АММ(*P*) – N_h, которая требуется для идентификации указанной АММ(*P*) с заданной доверительной вероятностью [14]. Для $n < 5$ максимальной эффективностью (минимальной оценкой *N*) обладают признаки из *M* ($N_M = \min N$), при $n = 5$ - $N_M = N_V = \min N$, а для $n > 5$ - $N_V = \min N$. При этом $\lim\limits_{n\to\infty} N_F = N_M$ при увеличении *n*.

Утверждение 3 [4]**.** При решении задачи идентификации АММ(*P*), $P \in \{T, Б\}$, на основе реализуемых ими ЦМ, если для величин ε_V и ε_W, определенных согласно **(**12) и (13), выполняется неравенство $\varepsilon_W > \varepsilon_V$, то $N_W < N_V$.

Следствие из утверждения 3 [4]**.** Если для ЭСМ P_1 и P_2 для $\forall i = \overline{1, n}$, $\exists\left(\left(p_{ij}^{(1)} = 0\right) \wedge \left(p_{ij}^{(2)} = 0\right)\right)$ и $\varepsilon_W > \varepsilon_V$, то для $q_{lz}(i)$ и $q'_{lz}(i, j)$, определенных согласно (12) и (13) справедливо:

$$\min_{l,z=1,4,\, l\neq z}\left(n^{-1}D\sum_{i=1}^{n}\sum_{j=1}^{n} q'_{lz}(i, j)\right) > \min_{l,\, z=1,4,\, l\neq z}\left(\sum_{i=1}^{n} q_{lz}(i)\right) \qquad (14)$$

Таким образом, для решения задачи идентификации АММ(*P*) целесообразно использовать признаки *W* вместо *V*, если выполняется условие вида (14).

С целью исследования эффективности предложенного численного метода многопараметрической идентификации АММ(*P*), где *P* принадлежат к априори заданным подклассам, проведен факторный анализ (ФА) группы признаков, позволяющих идентифицировать подкласс ЭСМ *P*. В данную группу входят признаки из множеств *F*, *M* и *V*. При использовании ФА получены результаты для ЦМ, сгенерированных на основе АММ(*P*), где *P* имеет размерность *n* = 5 (длина генерируемых ЦМ равна 700). Данные решения позволяют сгруппировать (по факторам) признаки из множеств *V, F* и *M*, поместив в каждую из групп те из них, которые сильно коррелированы между собой.

В п. 2.3 предложена модель идентификации конечных простых однородных ЦМ, сгенерированных на основе ЭСМ класса циклических индекса *r* (ЦСМ$_r$). Модель основана на вычислении функционалов от элементов циклической ЦМ и позволяет выявить свойство цикличности ЦМ, определив при этом длину цикла *r*, а также вероятность идентификации циклической ЦМ в зависимости от ее длины *N*.

ЦСМ$_r$ имеет период $r > 1$, а ее состояния подразделяются на *r* циклических классов – $\overline{S_1}$, $\overline{S_2}$, …, $\overline{S_r}$: ЦМ движется по циклическим классам в определенном порядке, возвращаясь в класс с начальным состоянием через *r* шагов (Романовский И.В., 1949), при этом $n_i = \left|\overline{S_i}\right|$, $i = \overline{1, r}$, $\sum_{i=1}^{r} n_i = n$. Обозначим *r* возможных групп вариантов последовательностей, генерируемых на основе ЦСМ$_r$ в виде:

$$\chi_K = \left\{s_{i_1}^k, s_{i2}^k, ..., s_{i_r}^k\right\}, \text{ где } s_{i_1}^k \in \overline{S_{l+k}}, l = \overline{1, r-k}, \text{ для } k > 0 : s_{i_{r-k+j}}^k \in \overline{S_j},\ j = \overline{1, k}. \qquad (15)$$

Характеристические признаки ЦСМ$_r$ c_r, определяемые индексом *r*, $r > 1$, позволяют определить признаки, представленные в виде функционалов на ЦМ – $\tilde{c}_r$. В зависимости от значения индекса *r* ЦСМ$_r$ выделены подклассы ЭСМ. Если ЦСМ$_r$ имеет индекс *r*, $r > 1$, то признак c_r, позволяющий распознать АММ вида (2), заданную ЦСМ$_r$, определяется выражением [18]:

$$c_r = \begin{cases} 1, \text{если} \underset{i=0}{\overset{r-1}{\vee}} \left(\underset{j=0}{\overset{m-1}{\&}} s_{t_{i+j+1}} \in \overline{S}_{(i+j) \bmod r+1} \right), m = \overline{1, r} \\ 0, \text{иначе} \end{cases}, \quad (16)$$

Признаки c_r, определенные согласно (16), являются теоретическими оценками величин $\tilde{c}_r$, вычисление которых производится путем подсчета циклически повторяющихся подмножеств состояний ЦМ определенной длины N [18]:

$$\tilde{c}_r = p_{ID} \cdot c_r, \quad (17)$$

где p_{ID} – минимальная вероятность безошибочного распознавания ЦМ длины N к заданному подклассу SC_r. Справедливо

Утверждение 4[18]. Минимальная вероятность безошибочного распознавания циклической ЦМ вида (1), $|S| = n$ и длины N, определяемой на основе ЦСМ$_r$, положительные элементы которой определены с дискретностью D, равна:

$$p_{ID} = 1 - \overline{p}^{(r)}_{\max} = 1 - \left(\frac{n-1}{n} D \right)^N. \quad (18)$$

Предложен численный метод идентификации принадлежности АММ(P) к подклассу SC_r на основе порождаемых ей ЦМ, включающий четыре этапа [5]:1) формирование множества объектов: ЦМ длины N; 2) вычисление значений признака вида (16) для каждого из $S(n,r)$ разбиений множества состояний ЦМ на подклассы, $r = \overline{2, n}$; 3) идентификация принадлежности объекта к заданному подклассу SC_r, $r = \overline{2, n}$; 4) расчет вероятности правильной идентификации объекта согласно (18). Корректность метода идентификации, предложенного в [2], проверена путем проведения исследования выборки циклических ЦМ, состоящей из 140 объектов. Для идентификации АММ, заданного на основе ЦСМ$_r$ при n=4 требуется от 7 до 14 элементов ЦМ, при n=7 требуется от 14 до 36 элементов ЦМ. Верхняя оценка длины увеличивается при увеличении n. В результате процедуры идентификации все объекты исследуемой выборки были идентифицированы с доверительной вероятностью 0,995.

Обозначим $A(\chi_k)$ – КДА, распознающий последовательность χ_k вида (15), либо ее первые m элементов, $k = \overline{0, (r-1)}$. Выход $A(\chi_k)$ определен значением булевой переменной $\underset{j=0}{\overset{m-1}{\&}} s_{t_{i+j+1}} \in \overline{S}_{(i+j) \bmod r+1}$, $m = \overline{1, r}$. Задача оценки сложности распознавания принадлежности АММ к подклассу SC_r сводится к определению количества автоматов типа $A(\chi_k)$, распознающих всевозможные варианты последовательностей χ_k, $k = \overline{0, (r-1)}$.

Утверждение 5 [18]. Оценки временной и емкостной сложности распознавания принадлежности АММ вида (2), заданной на основе ЦСМ$_r$, не приведенной к нормальному циклическому виду, к подклассу SC_r, равны, соответственно: T и $Q = r \cdot S(n,r)$, где $S(n,r)$ – число Стирлинга II рода.

В п. 2.4 предложен численный метод анализа внутренней структуры кластеров, выделяемых путем многопараметрической классификации множества объектов, определяемых различными группами признаков [17, 19]. Кластеры выделяются на основе различных групп признаков, при использовании как иерархических методов кластерного анализа, так и дивизивного метода k-средних [3]. Метод основан на минимизации суммарной выборочной дисперсии элементов относительно кластерных центров. Он представляет собой пошаговое нахождение кластерных центров и разбиение выборки объектов на кластеры до тех пор, пока заданный функционал, не перестанет уменьшаться. Метод предполагает оптимизацию разбиений на кластеры путем введения во множество кластеризуемых объектов, представителей каждого из предполагаемых кластеров с заданными характеристиками [17, 19]. Указанный численный метод актуален, в

частности, для решения задачи многопараметрического анализа и ранжирования предложений НИОКР [3].

В третьей главе предложены модификации [4, 5, 7, 20-23] известных модели и алгоритма «прямого-обратного хода» для решения задачи идентификации АММ (Рабинер Л., 1989). В п. 3.1 предложена модификация модели и алгоритма «прямого-обратного хода». В трактовке алгоритма в рамках скрытой марковской модели ставится задача вычисления вероятности того, что последовательность, наблюдаемая на выходе, соответствует заданной модели. Предложенный в [7, 20-22] модифицированный алгоритм отличается возможностью решать задачу идентификации АММ на основе порождаемых ей ЦМ и позволяет определить значения $P(\hat{S}(N) \mid AMM(P))$ – вероятности того, что $\hat{S}(N)$ сгенерирована на основе $AMM(P)$, где ЭСМ P размерности $n \times n$ принадлежит заданному подклассу и имеет максимальную энтропию.

В п. 3.2 предложен алгоритм идентификации конечных простых однородных ЦМ, сгенерированных на основе ЭСМ, на основе алгоритма «прямого-обратного» хода [20, 21, 23].

Пусть $\hat{S}_k(N)$ – ЦМ длины N, вида, аналогичного $\hat{S}(N)$, для которой существуют k моментов времени, $k < N$, когда состояния $s(t)$ скрыты от наблюдения. Для идентификации последовательности $\hat{S}(N) = s(1)s(2)...s(N)$ введем массивы переменных:

$$\alpha_t(i) = P(s(1)s(2)..s(t), s(t) = s_i \mid AMM(P)),\ t = \overline{1,N},\ i = \overline{1,n},$$

которые позволяют определить вероятность того, что АММ(P) к моменту времени t порождает последовательность $\hat{S}(t) = s(1)s(2)...s(t)$ и в момент времени t АММ(P) находится в состоянии s_i. Находим значения $\alpha_t(i)$, $t = \overline{1,N}$, $i = \overline{1,n}$, методом индукции по алгоритму, созданному на основе модификации алгоритма прямого-обратного хода для скрытых ЦМ [21]. Алгоритм включает три этапа: 1) инициализация: $\alpha_1(i) = \pi_0(i) \cdot z_i$, $i = \overline{1,n}$, $z_i = \begin{cases} 1: & s(t+1) = s_j \\ 0: & \textit{иначе} \end{cases}$; 2) индукция: $\alpha_{t+1}(j) = \left[\sum_{i=1}^{m} \alpha_t(i) \cdot p_{ij}\right] \cdot z_j$, $t = \overline{1,N-1}$, $j = \overline{1,n}$; 3) расчет $P(\hat{S}(N) \mid AMM(P)) = \alpha_N(s(N))$.

Для идентификации последовательности, k состояний которой скрыто от наблюдения – $\hat{S}_k(N)$, при выполнении этапа 2 вычисления значений $\alpha_{t+1}(i)$, $t = \overline{1,N-1}$, $i = \overline{1,n}$, имеет место выражение, основанное на целочисленном выражении суммы вида: $\alpha_{t+1}(j) = \left[\sum_{i=1}^{n} \alpha_t(i) \cdot p_{ij}\right] \cdot z'_j$, $z'_j = \begin{cases} 1: & s(t+1) - \text{скрыто} \\ z_j: & \text{иначе} \end{cases}$. Кроме того, если $s(N)$ скрыто от наблюдения, то вероятность $P(\hat{S}_k(N) \mid AMM(P)) = \sum_{i=1}^{n} \alpha_N(i)$.

В п. 3.3. предложена модификация алгоритма «прямого-обратного хода», который, аналогично алгоритму из п. 3.2, включает этапы 1) − 3). Этапы повторяются для каждой $P \in P_n(\text{ЦСМ}_r)$, т.е. $S(n,r)$ раз. Решена задача идентификации конечных простых однородных циклических ЦМ, часть элементов которой скрыта от наблюдения [23].

В п. 3.4 представлены алгоритмы вычисления признаков, служащих для идентификации как конечных простых однородных ЦМ, сгенерированных на основе простых стохастических матриц класса эргодических, так и циклических ЦМ. Приведен сравнительный анализ сложности данных алгоритмов [4, 21].

Алгоритмы идентификации с использованием рассматриваемых в главе 2 групп признаков реализованы на основе метода идентификации, включающего три этапа: 1) инициализация последовательностей $\hat{S}(N)$ по заданным начальным параметрам модели; 2) расчет признаков

идентификации AMM(P) по $\hat{S}(N)$; 3) расчет вероятности идентификации $P(\hat{S}(N) \mid AMM(P))$ для каждого заданного подкласса P.

Порядок оценки сложности вычисления частотных признаков для идентификации класса AMM(P) по реализациям $\hat{S}(N)$ в зависимости от выбора группы признаков определяются согласно следующим утверждениям:

Утверждение 6 [4]. Порядок оценки сложности вычисления частотных признаков для идентификации класса AMM(P) по реализациям $\hat{S}(N)$ равен $O(Nn)$.

Утверждение 7 [4]. Порядок оценки сложности вычисления признаков $\tilde{v}_g$, $\tilde{w}_g$ и $\tilde{m}_g$ для идентификации класса AMM(P) по реализациям $\hat{S}(N)$ равен $O(Nn^2)$.

Для предложенного модифицированного алгоритма справедлива

Теорема 2 [7]. Вычислительная сложность поиска значения $P(\hat{S}_k(N) \mid AMM(P))$, если $s(N)$ не скрыто от наблюдения, составляет $(N-1+k(n-1))\cdot(n-1)$ операций умножения и сложения; если $s(N)$ скрыто от наблюдения, то вычислительная сложность увеличивается на $(n-1)$ операцию сложения.

Следствие из теоремы 2 [4]. Вычислительная сложность поиска значения $P(\hat{S}(N) \mid AMM(P))$ имеет порядок $O(Nn)$.

Наличие ненаблюдаемых элементов в количестве, сопоставимом с длинной последовательности N, увеличивает порядок вычислительной сложности алгоритма $O(N \cdot n^2)$. Согласно сравнению алгоритмов идентификации, предложенные модифицированный алгоритм «прямого-обратного хода» и алгоритм на основе признаков $\tilde{f}_k$, являются достаточно эффективными по оценкам вычислительной сложности. Использование же функционалов на основе l-грамм, для $l = 2,3$ целесообразно при небольших значениях n.

Вычислительная сложность алгоритмов получения вероятности идентификации циклических ЦМ с помощью функционалов, сформированных на базе характеристического признака циклической СМ, равно $O(r \cdot n \cdot N \cdot S(n,r))$. Для модифицированного алгоритма «прямого-обратного хода» вычислительная сложность поиска значения $P(\hat{S}(N) \mid AMM(P))$, $P \in P_n(\text{ЦСМ}_r)$, имеет порядок $O\left(n \cdot N \cdot S(n,r)\right)$ [6]. Если $s(N)$ скрыто от наблюдения, то вычислительная сложность увеличивается на n операций сравнения.

Таким образом, модифицированный алгоритм «прямого-обратного хода» имеет меньшие по сравнению с алгоритмом, предложенным в 2.3, оценки вычислительной сложности, в меньшей степени зависимые от индекса идентифицируемой циклической ЦМ. Указанная задача является актуальной для решения широкого круга задач идентификации цепей Маркова, в том числе – частично скрытых от наблюдения [7, 20 - 23].

В четвертой главе описан разработанный программный комплекс, реализующий предложенные модели, методы и алгоритмы [24-27]. Структура программного комплекса представлена на рисунке.

Подсистема сбора данных для анализа состоит из двух модулей: модуль ввода и генерации классов ЭСМ и модуль получения на их основе ЦМ. Подсистема выбора объектов по параметрам позволяет организовать выборку из базы данных по параметрам модели. Модуль идентификации циклических ЦМ предназначен для идентификации AMM(P), определенных на основе ЦСМ, и основан на программной реализации методов и алгоритмов, описанных в п. 3.1. Также в подсистеме реализован расчетный модуль для вычисления вероятности отнесения к классу AMM(P).

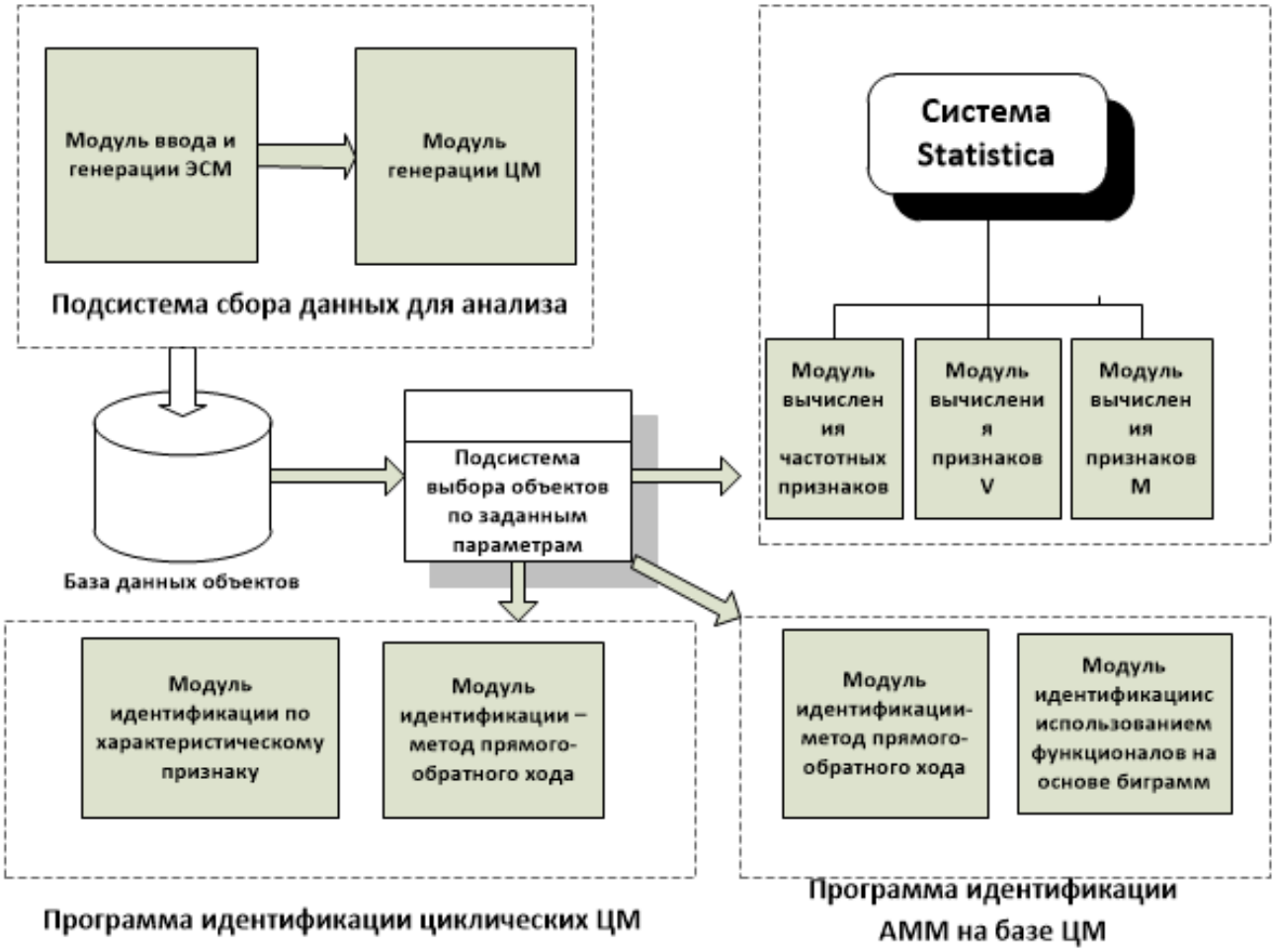

Рисунок − Структура программного комплекса

В модуле идентификации методом «прямого-обратного хода» реализованы алгоритмы, описанные в п. 3.2 и 3.3. Программный модуль идентификации АММ позволяет идентифицировать АММ(P) на основе ЦМ заданной длины, с возможностью вычисления доверительной вероятности корректной идентификации. Модуль реализован на основе методов и алгоритмов, описанных в пп. 2.3 и 3.4, соответственно. Подсистема классификации и идентификации АММ(P) основана на программной реализации методов и алгоритмов, описанных в п. 2.1 и 2.2. Она позволяет вычислить признаки по ЦМ и подготовить их для обработки в ИС Statistica 8.0. Программный комплекс реализован на языке Delphi с использованием интегрированной среды разработки ПО, Embarcadero® Delphi® XE3 Version 17.0.4625.53395.

В заключении сформулированы основные результаты диссертации.

ОСНОВНЫЕ РЕЗУЛЬТАТЫ РАБОТЫ

Разработан математический аппарат из области теории автоматов, теории распознавания и классификации, позволяющий повысить эффективность идентификации и классификации различных подклассов АММ на основе генерируемых дискретных цепей Маркова. Цель работы достигнута, получены следующие научные результаты.

1. Решена задача идентификации циклической ЦМ длины N и заданного индекса r на основе разработанных математической модели и модификации алгоритма «прямого-обратного хода». Автоматная марковская модель, реализующая ЦМ, задаваема на основе эргодической стохастической матрицы, не приведенной к нормальному циклическому виду. Критерии идентификации, характеристические признаки ЭСМ класса циклических, позволяют с заданной довери-

тельной вероятностью определить наличие циклов длины r в ЦМ длины N. Например, для $N \le 36$ и количества состояний ЦМ $n \le 7$ индекс ЦМ r определяется с доверительной вероятностью не менее 0,995. Предложенный модифицированный алгоритм «прямого-обратного хода», по сравнению с алгоритмом на основе вычисления функционалов от характеристических признаков ЦМ, имеет меньший порядок оценок вычислительной сложности – $O(n \cdot N \cdot S(n, r))$ и $O(r \cdot n \cdot N \cdot S(n, r))$.

2. Решена задача идентификации конечных простых однородных ЦМ к заданным подклассам на основе разработанных математической модели и модификации алгоритма «прямого-обратного хода». Подклассы ЦМ, генерируемых на основе АММ, определяемы подклассами ЭСМ, имеющих различную степень сходства-различия по структуре (квазитреугольные, нижние и верхние, и блочно-сообщающиеся, правые и левые) и задающих указанные АММ. Определен порядок вычислительной сложности предложенного алгоритма: если все элементы ЦМ наблюдаемы, то $O(N \cdot n)$; если же количество ненаблюдаемых элементов ЦМ сопоставимо с ее длиной N, то $O(N \cdot n^2)$.

3. Разработан подход к решению задачи многопараметрической классификации АММ на основе генерируемых ими ЦМ заданной длины N. АММ определяема ЭСМ, принадлежащими к заданным подклассам (см. п. 2). Согласно подходу из рабочего множества выбираются определенные признаки, позволяющие максимально эффективно классифицировать заданные АММ. Точность эмпирической классификации достигает 100% на выборке из 400 объектов − ЦМ с количеством состояний $n \le 5$ длины $N \le 564$. Увеличение количества состояний ЦМ способствует снижению длины N.

4. Предложен численный метод идентификации АММ, на основе генерируемых ими ЦМ с заданной доверительной вероятностью. АММ определена на основе априори заданных подклассов ЭСМ (см. п. 2). Множество признаков, вычисляемых по реализациям ЦМ с учетом структуры ЭСМ и точности представления ее элементов, повышает информативность решения задачи идентификации принадлежности генератора однородной ЦМ, задаваемого АММ вида (2), к заданному подклассу. Т.е. идентификация производится с большей доверительной вероятностью (по сравнению с предыдущими работами в данной области) для меньшего количества генерируемых элементов ЦМ − N. Для АММ с количеством состояний $n = 5$ идентификация ЦМ длины не более 700 выполнима с доверительной вероятностью не менее 0,95. Для $n < 5$ максимальной эффективностью (минимальной оценкой N) обладают признаки из M ($N_M = \min N$), при $n = 5$ - $N_M = N_V = \min N$, а для $n > 5$ - $N_V = \min N$. При этом $\lim_{n \to \infty} N_F = N_M$ при увеличении n. На основании результатов полученных факторных решений следует, что использование функционалов из группы V, частотных признаков множества F и признаков группы M является перспективным в плане повышения информативности объединенного множества признаков, используемого для идентификации АММ(P), $P \in \{T, Б\}$.

5. Разработан численный метод анализа внутренней структуры кластеров – дисперсии элементов каждого из кластеров относительно соответствующего кластерного центра; метод позволяет определить количество кластеров, оптимальное по предложенным критериям.

6. Создан комплекс программ, основу которого составляют связанные по входным данным алгоритмы и программы, служащие для реализации численных методов многопараметрического анализа и алгоритмов, предложенных в главах 2 и 3, соответственно. Проведены экспериментальные исследования предложенных математических моделей.

РАБОТЫ АВТОРА ПО ТЕМЕ ДИССЕРТАЦИИ

Публикации в ведущих рецензируемых научных изданиях.

1. Нурутдинова, А.Р. Методика идентификации автоматных марковских моделей на основе порождаемых ими последовательностей/ А.Р.Нурутдинова, С.В.Шалагин// Вестник КГТУ им. А.Н.Туполева. – 2010. – № 1. – С. 94 – 99.

2. Нурутдинова, А.Р. Многопараметрическая классификация автоматных марковских моделей на основе генерируемых ими последовательностей состояний/А.Р. Нурутдинова, С.В.Шалагин //Прикладная дискретная математика.– 2010. – № 4.– С. 41–54.

3. Барковский, С.С. Многопараметрический анализ и ранжирование предложений НИОКР отраслевой программы/ С.С.Барковский, А.Р.Нурутдинова, С.В.Шалагин// Вестник КГТУ им. А.Н.Туполева. – 2011. – № 2. – С. 115 – 122.

4. Шалагин, С.В. Сравнительный анализ вычислительной сложности алгоритмов идентификации конечных простых однородных цепей Маркова/ Шалагин С.В., Нурутдинова А.Р.// Вестник КГТУ. – 2016. – N 13. – С.153-156.

5. Nurutdinova, A.R. Identification algorithms of simple homogeneous Markov chains of cyclic class and their complexity analysis/ Nurutdinova A.R., Shalagin S.V.// International Journal of Pharmacy & Technology.−2016.-№3.−P. 18926-18935.

6. Nurutdinova, A.R. Identification of Markovian automata sub-classes/ Nurutdinova A.R., Shalagin S.V.// International Journal of Pharmacy & Technology. – 2016. – №3. – P. 15327-15337.

7. Нурутдинова, А.Р. Модифицированный алгоритм «прямого-обратного хода» решения задачи идентификации автоматных марковских моделей /А.Р. Нурутдинова // Системы управления и информационные технологии. – 2018. – № 2(72). – С. 36–41.

Публикации в сборниках трудов и материалов конференций и семинаров.

8. Сабитова, А.Р. Дискриминантный анализ вероятностных моделей марковского типа/ А.Р.Сабитова, С.В.Шалагин// Наука. Технологии. Инновации: Матер. всерос. научн. конф. молодых ученых. − Новосибирск: изд-во НГТУ, 2007. − Ч. 1. − С. 90 – 92.

9. Сабитова, А.Р. Многопараметрическая классификация марковских последовательностей/ А.Р.Сабитова, С.В.Шалагин// XV Туполевские чтения: межд. молодежная научная конф.: Тр. конф. − Казань: изд-во КГТУ, 2007. − С. 78 – 79.

10. Шалагин, С.В. Классификация стохастических матриц методом дискриминантного анализа/ С.В.Шалагин, А.Р.Сабитова// Наука и профессиональное образование: Матер. регион. НПК. – Нижнекамск: из-во КГТУ, 2007. − С. 217 – 220.

11. Сабитова, А.Р. Многопараметрическая классификация генераторов цепей Маркова/ А.Р.Сабитова, С.В.Шалагин// Проблемы техники и технологий телекоммуникации, ПТиТТ-2008: Тез. докл. 9-й межд. НТК. − Казань: изд-во КГТУ им. А.Н.Туполева, 2008. − С. 145 – 147.

12. Сабитова, А.Р. Вейвлет-анализ последовательностей, порождаемых автоматными марковскими моделями/ А.Р.Сабитова, С.В.Шалагин//XVI Туполевские чтения: межд. молодежная научная конф: Тр. конф. Т. III. − Казань: изд-во Казан. гос. техн. ун-та, 2008. − С. 78 – 79.

13. Шалагин, С.В. Исследование дискретных случайных процессов, реализуемых автоматными марковскими моделями/ С.В.Шалагин, А.Р.Сабитова// Наука: современное состояние и перспективы развития. Матер. всерос. научной конф. – Нижнекамск: из-во Казан. гос. техн. ун-та, 2009 . − С. 180 – 183.

14. Шалагин, С.В. Лингвистический подход к классификации последовательностей марковского типа/ С.В.Шалагин, А.Р.Нурутдинова// Инфокоммуникационные технологии глобального информационного общества: Сб. тр. 7-й межд. НПК. − Казань: изд-во Казан. гос. техн. ун-та, 2009. − С. 142 – 145.

15. Нурутдинова, А.Р. Идентификация автоматных марковских моделей на основе функционалов от биграмм/ А.Р.Нурутдинова, С.В.Шалагин// Информационные технологии в системе социально – экономической безопасности России и её регионов: Сб. тр. II всерос. научн. конф. − Казань: изд-во ТГГПУ, 2009. − С. 242 − 246.

16. Шалагин, С.В. Анализ признаков, идентифицирующих автоматные марковские модели по генерируемым ими последовательностям/ С.В.Шалагин, А.Р.Нурутдинова//Синтез и сложность управляющих систем: Матер. 18-й Междунар. школы-семинара им. акад. О.Б. Лупанова. – М: изд-во механико-математич. фак-та МГУ,2009. − С 109 − 113.
17. Шалагин, С.В. Методики кластеризации методом «*k*-средних» при использовании заданных критериев качества/ С.В.Шалагин, А.Р.Нурутдинова// Информационные технологии в профессиональной деятельности и научной работе: сб. матер. всерос. НПК с межд. уч. − Йошкар-Ола: МГТУ, 2010. Ч. 2. − С. 44 − 48.
18. Нурутдинова, А.Р. Распознавание подклассов марковских автоматов на основе последовательности состояний конечной длины/ А.Р.Нурутдинова, С.В.Шалагин// Проблемы теор. кибернетики: матер. XVI межд. конф. – Нижн. Новгород: изд-во ННГУ, 2011. − С. 344 – 347.
19. Нурутдинова, А.Р. Многопараметрический анализ автономных вероятностных автоматов/ А.Р.Нурутдинова, Р.Н. Шайхмурзина// XIX Туполевские чтения: матер. XIX межд. конф. Т. III. − Казань: изд-во Казан. гос. техн. ун-та, 2011. − С. 337 – 338.
20. Шалагин, С. В. Модель идентификации конечных простых однородных цепей Маркова/ С.В.Шалагин, А.Р.Нурутдинова // Актуальные проблемы и перспективы развития гражданской авиации России. Сб. тр. V НТК преподавателей, научных работников и аспирантов с междунар. участием. – Иркутск: Иркутский филиал МГТУ ГА, 2016. − С. 116 – 122.
21. Шалагин, С.В. Идентификация источника последовательности на основе скрытых цепей Маркова/ С.В.Шалагин, А.Р.Нурутдинова// Новые технологии, материалы и оборудование российской авиакосмической отрасли: Сб. докл. всерос. НПК с межд. участием. Т. 2. – Казань: Изд-во АН РТ, 2016. – С.292-297.
22. Шалагин, С.В. Метод решения задачи идентификации автоматных марковских моделей/ С.В.Шалагин, А.Р.Нурутдинова//Динамика нелинейных дискретных электротехнических и электронных схем: матер. 12-ой Всерос.науч.-техн. конф. – Чебоксары: Изд-во Чуваш.ун-та, 2017. – С. 12-14.
23. Шалагин, С.В. Идентификация последовательности измерений экономических параметров на основе скрытой марковской модели/С.В.Шалагин, А.Р.Нурутдинова// Проблемы анализа и моделирования региональных социально-экономических процессов: материалы докладов VII Межд. очн. науч.-практ. конф. – Казань: изд-во Казан. ун-та, 2017. – С. 159 – 162.

Свидетельства о регистрации программ для ЭВМ.

24. Свид. о гос. рег. программы для ЭВМ 2014662386 РФ. Идентификация марковских автоматов на основе порождаемых ими последовательностей состояний/ А.Р. Нурутдинова; заявитель и патентообладатель Нурутдинова А.Р. – № 2014660063; Заявл. 06.10.2014; Зарег. в реестре программ для ЭВМ 28.11.2014; Опубл. 20.12.2014, Бюл. № 12(98).
25. Свид. о гос. рег. программы для ЭВМ 2015615009 РФ. Распознавание подклассов марковских автоматов, определенных на базе циклических стохастических матриц / А.Р. Нурутдинова; заявитель и патентообладатель Нурутдинова А.Р. – № 2015611985; Заявл. 17.03.2015; Зарег. в реестре программ для ЭВМ 06.05.2015; Опубл. 20.06.2015, Бюл. № 6(104).
26. Свид. о гос. рег. программы для ЭВМ 2017663314 РФ. Программа распознавания цепей Маркова методом прямого-обратного хода/ А.Р. Нурутдинова, Д. Р. Григорьева; заявитель и патентообладатель ФГАОУ ВО КФУ – № 2017615601; Заявл. 13.06.2017; Зарег. в реестре программ для ЭВМ 28.11.2017; Опубл. 20.12.2017, Бюл. № 12.
27. Свид. о гос. рег. программы для ЭВМ 2016661744 РФ. Программа генерации цепей Маркова и эргодических стохастических матриц по заданным классам автоматной марковской модели / А.Р. Нурутдинова, Р.В. Валиев; заявитель и патентообладатель Нурутдинова А.Р. – № 2016619262; Заявл. 31.08.2016; Зарег. в реестре программ для ЭВМ 20.10.2016; Опубл. 20.11.2016 , Бюл. № 10.

* В 2011 году автор поменяла фамилию Сабитова на Нурутдинова. Свид. о гос. регистрации брака № I-КБ №845660 от 25 июля 2011г.