На правах рукописи

Гультяева Татьяна Александровна

ИССЛЕДОВАНИЕ ПОДХОДА К РЕШЕНИЮ ЗАДАЧИ КЛАССИФИКАЦИИ ПОСЛЕДОВАТЕЛЬНОСТЕЙ, ПРЕДСТАВЛЕННЫХ СКРЫТЫМИ МАРКОВСКИМИ МОДЕЛЯМИ, С ИСПОЛЬЗОВАНИЕМ ИНИЦИИРОВАННЫХ ЭТИМИ МОДЕЛЯМИ ПРИЗНАКОВ

Специальность 05.13.17 – Теоретические основы информатики

АВТОРЕФЕРАТ
диссертации на соискание ученой степени
кандидата технических наук

Новосибирск – 2013

Работа выполнена в Федеральном государственном бюджетном образовательном учреждении высшего профессионального образования «Новосибирский государственный технический университет»

Научный руководитель: доктор технических наук, профессор
Попов Александр Александрович

Официальные оппоненты: Загоруйко Николай Григорьевич
доктор технических наук, профессор
Федеральное государственное бюджетное учреждение науки «Институт математики им. С.Л. Соболева» Сибирского отделения Российской академии наук, заведующий лабораторией анализа данных;

Фаддеенков Андрей Владимирович
кандидат технических наук, доцент
Федеральное государственное бюджетное образовательное учреждении высшего профессионального образования «Новосибирский государственный технический университет», доцент кафедры теории рынка

Ведущая организация: Федеральное государственное бюджетное образовательное учреждение высшего профессионального образования «Томский государственный университет систем управления и радиоэлектроники»

Защита состоится « 26 » декабря 2013 г. в 14^{00} часов на заседании диссертационного совета Д 212.173.06 при Федеральном государственном бюджетном образовательном учреждении высшего профессионального образования «Новосибирский государственный технический университет» по адресу: 630073, Новосибирск, пр. К. Маркса, 20.

С диссертацией можно ознакомиться в библиотеке Новосибирского государственного технического университета.

Автореферат разослан «___» ноября 2013 г.

Ученый секретарь
диссертационного совета Чубич Владимир Михайлович

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы исследований. Распознавание образов является одной из задач анализа данных, лежащего в основе любых научных методов. Один из подходов к классификации образов состоит в построении математической модели, описывающей эти образы. В диссертационной работе рассматриваются скрытые марковские модели (СММ), относящиеся к классу статистических моделей. Особенностью СММ является то, что они учитывают внутреннюю структуру последовательностей, которые регистрируются при наблюдении за процессами или объектами. СММ, собственно как метод, был разработан в середине 60х–конце 70х годов 20 века независимо отечественными (Винцюк Т. К., Ковалевский В.А.) и зарубежными (Баум Л.Е, Петри Т.) исследователями. Значительный вклад в развитие как теории, так и практического использования СММ внесли такие отечественные ученые, как Моттль В.В. и Мучник И.Б., Воробьев С.А., Борисов А.В. Зарубежные исследования таких ученых, как Рабинер Л.Р., Нефиан А.В., Самариа Ф., Каппе О., более ориентированные на прикладные задачи, также внесли свой вклад в развитие СММ. По своей природе СММ позволяют учитывать как пространственные, так и временные характеристики последовательностей. Поэтому эти модели получили широкое применение в различных прикладных задачах, таких как распознавание речи (см., например, работы авторов Рабинер Л.Р., Леггеттер С.Ж., Жанг Б.Х., Галес М., Хуанг К.Д., Гефке Д.А., Иконин С.Ю.), изображений (Нефиан А.В., Самариа Ф., Плотц Т., Бунке Х., Сандерсон К., Ли Дж.), видеопоследовательностей (Кошал А., Лиу К., Саид У.), задачи биоинформатики (Коски Т., Барбю В.С., Бирней Е., Дурбин Р.), задачи геофизики (Грант Р.), эконометрические задачи (Мамон Р.С., Грегоир С., Бхар Р.), задачи сферы телекоммуникаций (Миллер Б.М., Антон-Харо С., Фоноллоза Ж.А.), исследование динамических систем (Фрейзер А.М.), стегоанализ (Сидоров М.А.) и т. д.

Классификация последовательностей с использованием СММ при условии того, что конкурирующие модели достаточно хорошо различимы друг от друга (по вероятности), как правило, не вызывает затруднений. При этом традиционно используется подход (ТП – традиционный подход), основывающийся на критерии правдоподобия. Однако в случае, когда эти модели оказываются по какой-либо причине близкими, результаты классификации становятся малоинформативными, т. е. принадлежность некоторой последовательности к любому классу является равновероятной. Существует два подхода к решению этой задачи. В первом подходе используются методы, которые ориентированы на то, чтобы точнее описать исследуемый объект или процесс: либо изменяют структуру используемых моделей (см., например, работы таких авторов, как Александров В., Нил Р.М.), либо используют иные методы оценки параметров моделей (Вальдер С.Ж., Икбал С., Жоу Г.Д.), либо же комбинируют эти два подхода (Лиу С., Чатзис П.С.). При использовании второго подхода появляется возможность видоизменять решающее правило классификации, получая некую информацию от моделей и применяя ее для проведения классификации (см., например, работы таких авторов, как Солера-Урена Р., Линг Ч., Аран О.). Этот под-

ход представляется нам наиболее перспективным, т. к. здесь существует определенная свобода как в выборе пространства признаков, в котором проводится классификация, так и в выборе самого классификатора. Однако систематических исследований этого подхода в литературе не встречается.

Цель и задачи исследований. Основной целью диссертационной работы является исследование проблемы повышения дискриминирующих свойств СММ с помощью подхода, основанного на проведении классификации в пространстве первых производных от логарифма функции правдоподобия по параметрам СММ.

Для достижения поставленной цели предусмотрено решение следующих задач:

- исследование возможностей использования пространства признаков в виде производных от логарифма функции правдоподобия по параметрам СММ для классификации последовательностей;
- исследование поведения классификаторов, использующих пространство признаков в виде производных от логарифма функции правдоподобия по параметрам СММ в условиях, когда последовательности искажены помехами или при структурной неопределенности;
- сравнительный анализ методик выбора информативных признаков для построения пространства, в котором последовательности различаются наилучшим образом;
- выявление ситуаций, когда исследуемый подход (ИП) дает результаты лучше, чем традиционный подход, используемый для классификации последовательностей.

Методы исследования. Для решения поставленных задач использовался аппарат теории вероятностей, математической статистики, вычислительной математики, математического программирования, статистического моделирования.

Научная новизна диссертационной работы заключается в том, что были впервые:

- проведены исследования с применением технологии статистического моделирования, показавшие возможность решения задачи классификации порожденных скрытыми марковскими моделями последовательностей с использованием инициированных этими моделями признаков;
- получены выражения для вычисления производных от логарифма функции правдоподобия по параметрам скрытых марковских моделей, учитывающие используемое при работе с длинными последовательностями масштабирование;
- проведен сравнительный анализ ИП и ТП к решению задачи классификации порожденных скрытыми марковскими моделями последовательностей, искаженных помехами;
- проведен сравнительный анализ ИП и ТП к решению задачи классификации порожденных скрытыми марковскими моделями последовательностей в условиях неточных знаний о структуре этих моделей;

– выявлены особенности, присущие классифицируемым последовательностям, при которых использование ИП предпочтительнее ТП.

Основные положения, выносимые на защиту:

– результаты исследования ИП в условиях действия помех различной интенсивности и природы;
– результаты исследования ИП в условиях различных отклонений от традиционных предположений, налагаемых на последовательности и процессы, их порождающие;
– результаты исследований различных методик выбора информативных признаков для классификации последовательностей;
– выявленные особенности последовательностей, при которых использование ИП предпочтительнее ТП.

Обоснованность и достоверность научных положений, выводов и рекомендаций обеспечивается:

– корректным применением аналитического аппарата математического анализа, теории вероятностей и математической статистики для исследования свойств построенных моделей;
– подтверждением аналитических выводов и рекомендаций результатами статистического моделирования.

Личный творческий вклад автора заключается в проведении исследований, обосновывающих основные положения, выносимых на защиту.

Практическая ценность результатов заключается в том, что:

– результаты проведенных исследований позволяют корректно применять ИП в случаях, когда конкурирующие скрытые марковские модели близки по параметрам, последовательности искажены помехами или в условиях неточных знаний о структуре скрытых марковских моделей;
– проведены исследования, которые выявили особенности, присущие анализируемым последовательностям, при которых использование ИП дает выигрыш в сравнении с ТП;
– разработана программная система, позволяющая классифицировать последовательности, применяя ИП и ТП.

Апробация работы. Основные результаты исследований, проведенных автором, докладывались и обсуждались на: Российской НТК «Информатика и проблемы телекоммуникаций» (Новосибирск, 2010, 2011, 2012); Пятнадцатой всероссийской конференции «Математические методы распознавания образов-15» (Петрозаводск, 2011); Международной заочной научной конференции «Технические науки: проблемы и перспективы» (Санкт-Петербург, 2011); Всероссийской научной конференции молодых ученых «Наука. Технологии. Инновации» (Новосибирск, 2011); Четвертой международной конференции по распознавания образов и искусственному интеллекту «PReMI-2011» (Москва, 2011); XI международной НТК «Актуальные проблемы электронного приборостроения АПЭП-2012» (Новосибирск, 2012); Всероссийской конференции с международным участием «Информационные и математические технологии в науке, технике, медицине» (Томск, 2012).

Реализация полученных результатов. Результаты диссертационных исследований использованы при разработке автономной системы электроснабжения летательных аппаратов в рамках НИОКР с предприятием ОАО «АКБ Якорь», г. Москва, что подтверждено соответствующей справкой об использовании результатов диссертационной работы. Работа выполнялась при финансовой поддержке: по гранту № 2.1.1/2932 «Оптимизационные методы построения логико-вероятностных моделей в задачах многомерного статистического анализа» федеральной целевой программы «Развитие научного потенциала высшей школы» (2009-2010 гг.); по гранту № 2.1.1/11599 «Методы устойчивой идентификации в задачах построения зависимостей и классификаций по неоднородным экспериментальным данным в условиях параметрической и структурной неопределенности» федеральной целевой программы (2011 г.); стипендии Правительства Российской Федерации на 2011-2012 учебный год согласно приказу № 154 от 28.02.2012.

Публикации. Основные научные результаты диссертации опубликованы в 18 печатных работах, из них 5 – 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, 4 – в сборниках научных работ, 9 – в материалах конференций.

Структура работы. Диссертация состоит из введения, 5 глав, заключения, 6 приложений и списка использованных источников (212 наименований). Основной текст работы изложен на 242 страницах, включает 7 таблиц и 93 рисунка.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

Во введении обоснована актуальность темы, сформулированы цели и задачи работы.

В первой главе приводится постановка задачи исследования.

В п. 1.1 осуществляется введение в теорию СММ. Такие модели являются частным случаем скрытого марковского процесса, представляющего собой двухкомпонентный случайный процесс со скрытой компонентой (марковская цепь с конечным множеством состояний) и наблюдаемой компонентой. СММ задают следующие параметры:

1) вектор вероятностей начальных состояний $\Pi = \{\pi_i\}$, $i = \overline{1, N}$, где $\pi_i = P\{q_1 = s_i\}$, множество скрытых состояний $S = \{s_1, s_2, ..., s_N\}$, N – количество скрытых состояний в модели, q_1 – состояние в момент $t = 1$ из последовательности Q, являющейся реализацией скрытого процесса;

2) матрица вероятностей переходов $A = \{a_{ij}\}$, $i, j = \overline{1, N}$, где $a_{ij} = P\{q_t = s_j \,|\, q_{t-1} = s_i\}$, q_t – состояние в момент $t = \overline{2, T}$, T – длина последовательности;

3) функции условной плотности распределений наблюдений $B=\{b_i(t)\}$, где $b_i(t)$ – это условные плотности вероятностей $P\{o_t|q_t=s_i\}$, $o_t\in\mathbb{R}$ – наблюдение, фиксируемое в момент $t=\overline{1,T}$, из последовательности O, являющейся реализацией наблюдаемого процесса. В работе рассматривается случай, когда условные плотности распределений наблюдений являются смесями нормальных распределений:

$$b_i(t)=\sum_{m=1}^{M}\tau_{im}\left(\sqrt{2\pi}\sigma_{im}\right)^{-1}e^{-(o_t-\mu_{im})^2/2\sigma_{im}^2}, \qquad (1)$$

где τ_{im} – это вес m-ой компоненты смеси в описании i-ого скрытого состояния, M – количество компонент смеси, параметры μ_{im} и σ_{im}^2 являются, соответственно, математическим ожиданием и дисперсией.

В п. 1.2 рассматривается вопрос обучения СММ по имеющимся последовательностям исследуемого процесса или объекта. В работе был использован алгоритм обучения, в общем случае называемый EM (EM – Expectation Maximization; максимизация ожидания) или, применительно к СММ, алгоритмом Баума-Велша. Чтобы решить проблему того, что результаты умножений различных вероятностей, используемых для оценивания параметров СММ, стремятся к нулю, вводится масштабный коэффициент.

В п. 1.3 описывается методика моделирования последовательностей, используемая при проведении вычислительных экспериментов.

В п. 1.4 рассматривается ТП к классификации последовательностей.

В п. 1.5 приводится исторический обзор теории СММ, анализируются основные существующие на данный момент проблемы классификации последовательностей, описываемых СММ. Рассматриваются различные подходы к повышению дискриминирующих способностей СММ.

Во второй главе рассматривается задача классификации последовательностей.

В п. 2.1 приводится постановка задачи классификации последовательностей с принятием решений по прецедентам. При этом используются гипотезы о компактности и монотонности пространства решений.

В п. 2.2 описывается ИП к решению задачи классификации последовательностей. Для каждой обучающей последовательности O формируется характеристический вектор вида: $V=\left[Z(O,\lambda_1)\quad Z(O,\lambda_2)\right]^T$, где $Z(O,\lambda_k)$ – вектор первых производных от логарифма функции правдоподобия последовательности O по параметрам моделей λ_k, $k=\overline{1,2}$. Аналогичным образом вычисляется характеристический вектор для тестовой последовательности. Далее в полученном пространстве признаков решается задача классификации.

В п. 2.3 и п. 2.4 описываются применяемый в работе классификатор k ближайших соседей (kNN – k Nearest Neighbor) и классификатор, основанный на методе опорных векторов (SVM – Support Vector Machines).

В п. 2.5 рассматриваются известные способы многоклассовой классификации, сводящиеся к двухклассовой классификации.

В п. 2.6 приводятся формулы для вычисления первых производных от логарифма функции правдоподобия по различным параметрам СММ.

В п. 2.7 описываются особенности вычисления первых производных от логарифма функции правдоподобия для СММ при длинных последовательностях с учетом масштабного коэффициента.

В п. 2.8 приводятся результаты исследования возможности проведения классификации в пространстве первых производных по параметрам моделей для ИП в случае, если две конкурирующие модели имеют следующие параметры: $A_{ii}^{\lambda 2} = A_{ii}^{\lambda 1} + d^A$ при $i = \overline{1, N}$ и $A_{ii+1}^{\lambda 2} = A_{ii+1}^{\lambda 1} - d^A$ при $i = \overline{1, N-1}$, $A_{N1}^{\lambda 2} = A_{N1}^{\lambda 1} - d^A$, $\mu_{im}^{\lambda 2} = \mu_{im}^{\lambda 1} + d^\mu$, $\sigma_{im}^{\lambda 2} = \sigma_{im}^{\lambda 1} + d^\sigma$, $m = \overline{1, M}$. Таким образом, параметры d^A, d^μ, d^σ определяют степень близости между моделями. На рисунке 1 приведены зависимости среднего процента верно классифицированных последовательностей (СПВКП) для ситуации, когда модели отличались друг от друга только в матрицах переходных вероятностей и использовался ИП с SVM классификатором*. Данное исследование показало, что приемлемый результат по точности классификации достигается только в том случае, если используется пространство производных по тем параметрам, по которым при моделировании было заложено отличие между двумя конкурирующими моделями. При этом при увеличении длины последовательностей результаты классификаторов приближаются к результатам ТП на основе отношения функций правдоподобия. Таким образом, проведенные вычислительные эксперименты подтверждают выдвинутое предположение о том, что возможно использовать ИП с kNN и SVM классификаторами для классификации последовательностей в пространстве инициированных признаков – первых производных от логарифма функции правдоподобия по параметрам модели. Кроме того было установлено, что результаты классификации, полученные с использованием SVM классификатора в рассматриваемом подходе, несколько лучше результатов, полученных с использованием классификатора kNN.

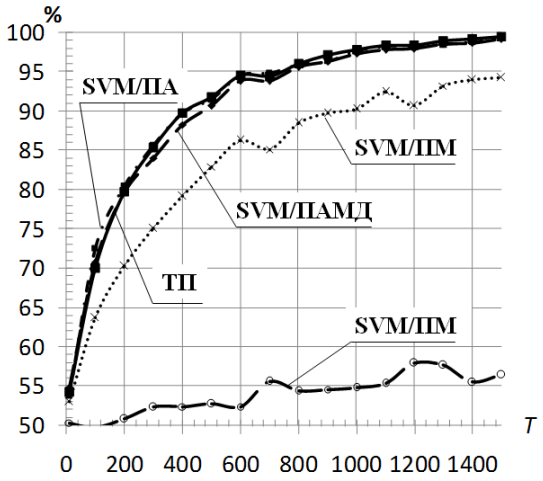


Рисунок 1 – Зависимость СПВКП от длины Т, $d^A = 0.1$

В третьей главе приведены результаты исследования поведения классификаторов в случае, когда наблюдаемые последовательности искажены вслед-

* На всех рисунках, приведенных в автореферате, используются следующие аббревиатуры для обозначения пространств производных от функции правдоподобия по: элементам матрицы A – ПА; параметрам математического ожидания – ПМ; параметрам дисперсии – ПД и пространство, состоящее из этих всех объединенных пространств – ПАМД.

ствие действия на них различных помех.

В п. 3.1 описываются используемые модели помех, действующих на последовательности. В работе используется искажение наблюдаемых последовательностей аддитивным шумом e с весом ω, показывающим насколько сильно этот шум искажает последовательность, и частичное замещение последовательности помехой по вероятностной схеме.

В п. 3.2 приводятся результаты исследований в случае, когда вероятности появления наблюдений описываются одним распределением. При распределении аддитивной помехи, подчиняющейся нормальному закону с постоянными параметрами распределения, классификатор kNN в пространстве ПА не дает улучшений в сравнении с ТП. Это связано с тем, что у шума и исходной последовательности вероятности появления наблюдений имеют нормальное распределение. Поэтому при проверке сложной гипотезы о согласии эмпирического распределения частот появления наблюдений в скрытых состояниях с нормальным распределением с параметрами, оцененными по алгоритму Баум-Велша, наблюдался высокий достигнутый уровень значимости. Кроме того отметим, что с увеличением уровня шума модели становятся плохо различимы, т. к. расстояние между ними стремится к нулю из-за того, что распределения наблюдений в скрытых состояниях стремятся к одному и тому же распределению (распределению ошибки) для двух моделей. В случае распределения ошибки по закону Коши и при проверке сложной гипотезы о согласии эмпирических распределений частот появления наблюдений в скрытых состояниях с нормальным распределением, было установлено, что небольшое увеличение уровня шума приводит к тому, что нормальное распределение плохо описывает эти эмпирические распределения частот. При этом до уровня шума $\omega \le 0.3$ модели остаются еще достаточно далеки друг от друга по расстоянию, и ИП показывает на этом интервале лучшую классификацию, чем традиционный. Таким образом, ИП позволяет повысить процент верно классифицированных последовательностей в сравнении с ТП в случае, когда расстояние между моделями еще достаточно велико, но согласие эмпирического распределения частот появления наблюдений в скрытых состояниях с нормальным распределением не достигается. На рисунке 2 приведены зависимости СПВКП от уровня шума для случая, когда помеха носит аддитивный характер с изменяемыми параметрами распределения, которые зависят от номера скрытого состояния модели: $e_1 \succ C(0,0.1)$, $e_2 \succ C(5,0.1)$, $e_3 \succ C(10,0.1)$. Заметно существенное преимущество ИП перед ТП. При вероятностной помехе наблюдаются аналогичные результаты. В целом можно сделать вывод, что для близких по параметрам моделей с количеством компонент смеси

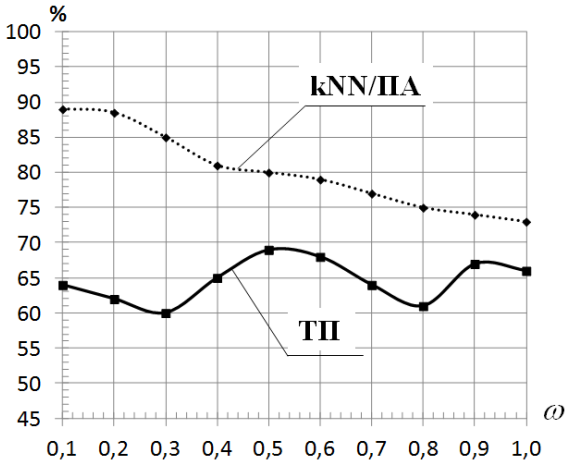


Рисунок 2 – Зависимость СПВКП от уровня шума ω. $d^A = 0.1$

$M=1$ при использовании классификатора kNN удается повысить процент верно классифицированных последовательностей в случае, если эти последовательности искажены помехой, имеющей распределение Коши. При этом прирост процентов верной классификации в сравнении с ТП может достигать 40%.

В п. 3.3 приводятся результаты исследований в случае, когда у СММ вероятности появления наблюдений описываются смесями распределений с количеством компонент смеси $M=3$. На рисунке 3 приведены зависимости СПВКП от уровня шума для случая, когда помеха носит аддитивный характер с постоянными параметрами распределения: $e \succ C(0,0.1)$. Заметно существенное преимущество ИП перед ТП. Аналогичная картина наблюдается и при различии между моделями, заложенными в d^{μ} и d^{σ}. Также в результате исследований было установлено, что ТП обладает наибольшим значением разброса СПВКП среди рассматриваемых методов классификации. Наименьшим разбросом усредненного процента обладает ИП, использующий SVM. Кроме того, почти всегда наблюдается ситуация, когда разброс усредненного процента для SVM меньше, чем разброс усредненного процента для kNN. При вероятностной помехе наблюдаются результаты аналогичные тем, что были получены при аддитивной помехе. Результаты сравнения поведения классификаторов на задаче двухклассовой классификации в целом остаются справедливыми и для многоклассового случая.

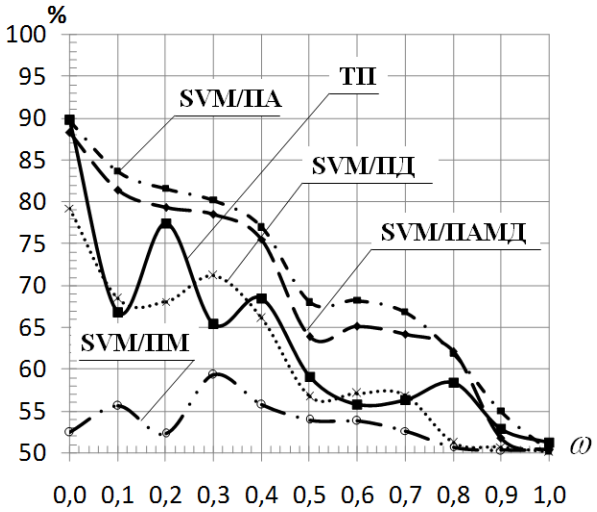


Рисунок 3 – Зависимость СПВКП от уровня шума ω. $d^{A}=0.1$

В п. 3.4 исследуются поведения классификаторов в условиях, когда помеха, смоделированная по вероятностной схеме, имеет различные законы распределения: Лапласа; минимального значения; двойного экспоненциального; логистического; Эрланга или же функцию вероятности распределения Пуассона. По результатам исследований, можно выдвинуть гипотезу о том, что ТП проигрывает в сравнении с ИП в случае, если последовательности искажены помехами с различным типом распределений: с тяжелыми хвостами, симметричными и асимметричными. Процент верно классифицированных последовательностей удается повысить всегда в сравнении с ТП, правильно выбрав пространство для классификации (в среднем на 10%).

В четвертой главе приведены результаты исследований в случаях, когда нарушены одни из априорных представлений либо о наблюдаемых последовательностях, либо о структуре СММ.

В п. 4.1 приводятся результаты исследований, когда используются различные законы распределений появления наблюдений. Рассматриваются различные случая отклонения от классического предположения о виде функции условной плотности распределений наблюдений (см. формулу (1)). На рисун-

ке 4 приведены результаты классификации для одного из случаев, когда эти функции для двух конкурирующих моделей выглядят следующим образом: $b_1^{\lambda_1}(t) = f_C(o_t;0,0.01)$, $b_1^{\lambda_2}(t) = f_C(o_t;0,0.01+d)$, $b_2(t) = f_C(o_t;0.05,0.01)$, $b_3(t) = f_C(o_t;-0.05,0.01)$, где f_C – функция плотности вероятности распределения Коши. Оценивание параметров моделей проводилось для СММ с количеством скрытых состояний и компонент смеси $N_{learn} = M_{learn} = 3$. Видно, что чем дальше модели раздвинуты друг от друга за счет увеличения значения параметра d, тем выше процент верной классификации в используемых пространствах. Кроме того, при $d \geq 0.1$ преимущество использования ИП теряется, т. к. модель смесей, которая используется при оценивании параметров, достаточно хорошо описала выбранный при моделировании закон распределения. После оценивания параметров можно было увидеть, что основное различие между двумя моделями оказалось сосредоточено в матрице переходных вероятностей, поэтому при классификации в пространстве ПА и в пространстве ПАМД наблюдается заметный выигрыш в сравнении ТП.

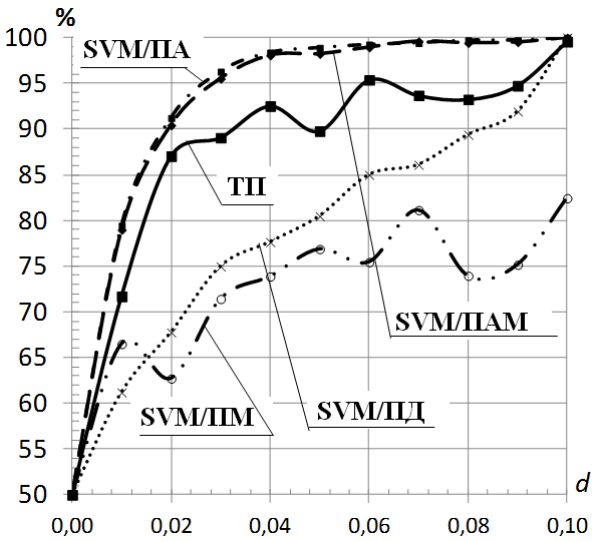


Рисунок 4 – Зависимость СПВКП от значения параметра d, $T = 100$

Были рассмотрены также случаи, когда распределение наблюдений подчинялось дискретному распределению вероятностей появления наблюдений (закон Пуассона), при этом было заметно ощутимое преимущество ИП. Также был использован закон Эрланга, однако выигрыша, как и проигрыша, перед ТП не наблюдалось, что объясняется тем, что смесь нормальных распределений достаточно хорошо аппроксимирует это распределение. Аналогичные результаты по преимуществу ИП перед ТП наблюдались и для остальных рассматриваемых видов функции условной плотности распределений наблюдений. По исследованиям, приведенным в этом параграфе, можно заключить, что даже если модель смеси нормальных распределений (1) недостаточно точно описывает истинное распределение наблюдений, то в целом СММ хорошо улавливает отличия между близкими моделями. Это подтверждается достаточно высоким средним процентом верной классификации при использовании ИП.

В п. 4.2 рассмотрен случай, когда исследователь предполагает, что у процесса или объекта, порождающего последовательности, нет скрытых состояний. Для исследования такой ситуации наблюдаемые последовательности для двух классов λ_1 и λ_2 моделировались по следующим формулам: $o_t^{\lambda_1} = \sin(0.1t) + e$, $o_t^{\lambda_2} = \left(1 + d^{Amp}\right)\sin\left(\left(0.1 + d^{\varpi}\right)t + d^{\varphi}\right) + e$, $e \succ N(0,0.1)$, $t = \overline{1,T}$; d^{Amp}, d^{ϖ}, d^{φ} – параметры, отвечающие за близость последовательностей. На рисунке 5 приведены зависимости СПВКП с использованием ТП и ИП с классификатором

SVM от количества скрытых состояний N_{learn} и компонент смеси M_{learn}, используемых на этапе обучения СММ. Наблюдается существенное преимущество во всех пространствах ИП в сравнении с ТП. Также проводились исследования при варьировании параметров близости последовательностей d^{Amp} и d^{φ}, с получением аналогичных результатов.

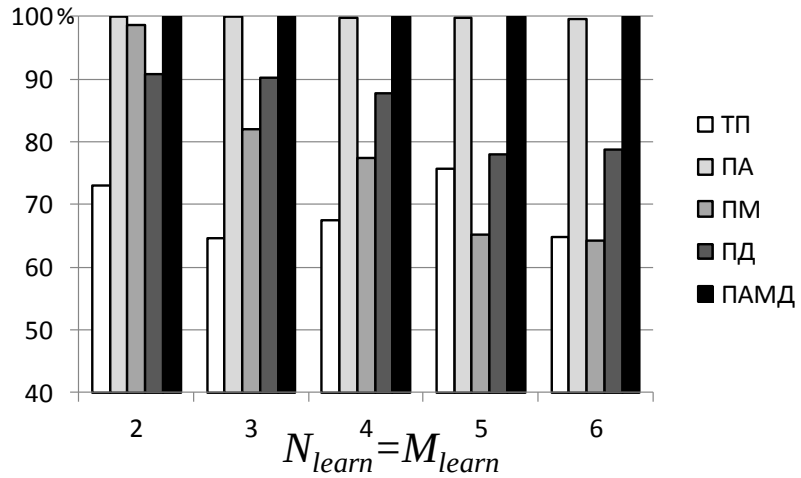
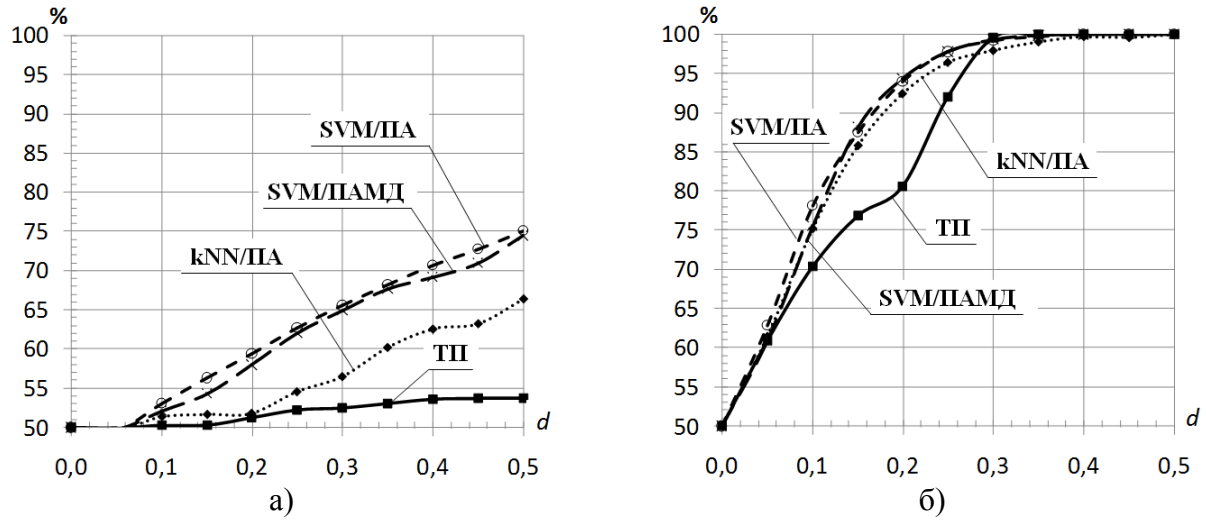


Рисунок 5 – Зависимость СПВКП от значения параметров N_{learn} и M_{learn}. $T=100$ для SVM, $d^{\varpi}=0.01$

В п. 4.3 исследуется поведение классификаторов в условиях структурной неопределенности. Исследования проводились для двухклассового случая при условии, что исследователь не обладает информацией об истинном значении параметров N и M, а конкурирующие модели близки друг к другу по элементам матрицы A. На рисунке 6 приведены зависимости СПВКП от параметра близостей моделей $d=d^{A}$ при количестве скрытых состояний и компонент смеси, с которыми моделировались последовательности $N=4$, $M=6$ и при различных значениях $N_{learn}=M_{learn}$ на этапах обучения и тестирования.



а) б)

Рисунок 6 – Зависимость СПВКП от параметра близостей моделей d при $N_{learn}=M_{learn}=2$ (а); $N_{learn}=M_{learn}=6$ (б)

Заметно, что при обучении моделей и тестировании возникает ситуация нехватки количества состояний для описания всей изменчивости смоделированной последовательности. ИП менее чувствителен к недобору количества как скрытых состояний, так и числа смесей, чем ТП. Для многоклассового случая получаются аналогичные выводы.

В пятой главе рассматривается решение задачи выбора информативных признаков и приводятся результаты ее решения для ряда модельных, экспериментальных и прикладной задач.

В п. 5.1 рассматриваются различные методы выбора информативных признаков. Как было установлено в работе, выбор пространства признаков, которое используется при классификации с применением ИП, существенным образом влияет на эффективность работы классификаторов. Оценка информативности признаков делается прямым способом с использованием средней ошибки распознавания, получаемой в режиме скользящего экзамена и одним из косвенных способов, основанным на вычислении значения функции конкурентного сходства (FRiS – Function of Rival Similarity).

В п. 5.2 приводятся результаты исследований на смоделированных данных. Можно говорить о том, что использование прямого подхода к выбору подсистемы признаков позволяет более точно выбрать пространство признаков, в котором будет наблюдаться наибольший процент верно классифицированных последовательностей на тестовых наборах. Использование косвенного похода на основе FRIS дает в целом неплохие результаты. Его можно рекомендовать к использованию в задачах с большой размерностью, требующих значительных вычислительных затрат.

В п. 5.3 проводятся исследования на широко распространенных выборках данных: «Hill-Valley Dataset» и «Waveform Database». В качестве прикладной задачи рассматривается определение по сигналу электрического тока структуры схемы электрической цепи, с которой был снят этот сигнал. Во всех случаях показано преимущество использования ИП в оптимально выбранном пространстве признаков перед ТП.

ЗАКЛЮЧЕНИЕ

В соответствии с поставленными задачами исследований получены следующие основные результаты.

1. Установлена возможность использования подхода к классификации представляемых СММ последовательностей, который основан на использовании признакового пространства в виде первых производных от логарифма функции правдоподобия по параметрам СММ.

2. Получены выражения для вычисления производных от логарифма функции правдоподобия по параметрам СММ, учитывающие используемое при работе с длинными последовательностями масштабирование.

3. Проведен сравнительный анализ ИП и основывающегося на критерии правдоподобия ТП к решению задачи классификации порожденных СММ последовательностей, искаженных помехами. При этом рассмотрены несколько вариантов распределений вероятностей появления помехи по непрерывным и дискретным законам. Установлено, что использование ИП позволят в ряде случаев повысить точность классификации на 30% в сравнении с ТП.

4. Проведен сравнительный анализ ИП и ТП к решению задачи классификации порожденных СММ последовательностей в условиях неточных знаний о

структуре этих моделей. А именно, рассмотрены различные законы распределений появления наблюдений, отличные от нормального. Установлено, что в этом случае использование ИП позволят повысить точность классификации на 45% в сравнении с ТП. Рассмотрена ситуация, когда процесс или объект, порождающий наблюдаемые последовательности, предположительно не имеет скрытых состояний. Установлено, что использование СММ для описания таких последовательностей и их классификации с помощью ИП позволят повысить точность классификации на 35% в сравнении с ТП. Рассмотрена ситуация, когда отсутствует априорная информация о количестве скрытых состояний и компонент смесей СММ. Установлено, что при заниженной оценке этих значений при использовании ИП можно повысить точность классификации на 40% в сравнении с ТП.

5. Проведено экспериментальное исследование методов выбора информативного подпространства признаков при использовании их в ИП по классификации последовательностей. Установлено, что выбор подпространства существенно влияет на точность классификации.

6. Выявлены особенности, присущие анализируемым последовательностям, при которых использование ИП предпочтительнее ТП. Эти особенности во многом определяют затрудненные условия применения ТП и связаны с близостью конкурирующих СММ, порождающих последовательности, и наличием различного рода искажений самих последовательностей.

7. Установлено, что разброс показателя точности классификации с использованием ТП выше, чем с использованием ИП в оптимально выбранном пространстве признаков.

8. ИП был применен для решения задачи классификации на ряде приводимых в литературе выборках, которые используются для оценки качества работы классификаторов. Установлено, что применяя ИП, удается повысить точность классификации в сравнении с результатами, приводимых другими исследователями для этих выборок. ИП был использован при разработке автономной системы электроснабжения летательных аппаратов в рамках НИОКР с предприятием ОАО «АКБ Якорь» г. Москва.

9. Разработана программная система, предназначенная для классификации одномерных последовательностей, тип элементов которых является числовым. Разработанное программное обеспечение используется при проведении научных исследований.

СПИСОК ПУБЛИКАЦИЙ ПО ТЕМЕ ДИССЕРТАЦИИ

Научные публикации в изданиях, входящих в
Перечень ведущих рецензируемых научных журналов и изданий:

1. Гультяева Т.А. Классификация последовательностей с использованием скрытых марковских моделей в условиях неточного задания их структуры / Т.А. Гультяева, А.А. Попов // Вестник ТГУ. Управление, вычислительная техника и информатика – Томск : Изд-во ТГУ, 2013. – № 3 (24). – С. 57-63.

2. Гультяева Т.А. Классификация смоделированных скрытыми марковскими моделями последовательностей в многоклассовом случае / Т.А. Гультяева, А.А. Попов // Научный вестник НГТУ. – Новосибирск : Изд-во НГТУ, 2013. – № 3 (52). – С. 40-45.

3. Гультяева Т.А. Классификация последовательностей, смоделированных скрытыми марковскими моделями при наличии аддитивного шума / Т.А. Гультяева, А.А. Попов // Научный вестник НГТУ. – Новосибирск : Изд-во НГТУ, 2012. – № 3 (48). – С. 17-24.

4. Гультяева Т.А. Классификация зашумленных последовательностей, порожденных близкими скрытыми марковскими моделями / Т.А. Гультяева, А.А. Попов // Научный вестник НГТУ. – Новосибирск : Изд-во НГТУ, 2011. – № 3 (44). – С. 3-16.

5. Gultyaeva T.A. The Classification of Noisy Sequences Generated by Similar HMMs / T.A. Gultyaeva, A.A. Popov // PReMI 2011, LNCS. – Springer-Verlag Berlin Heidelberg, 2011. – Vol. 6744/2011. – P. 30-35. [Классификация зашумленных последовательностей, смоделированных близкими СММ].

Научные публикации в других изданиях:

6. Гультяева Т.А. Исследование поведения классификаторов при классификации смоделированных СММ последовательностей в условиях отклонения от классических предположений / Т.А. Гультяева, А.А. Попов // Информационные и математические технологии в науке, технике, медицине : материалы Всерос. конф. с междунар. участием, Томск, 2012 г. – Томск : Изд-во ТПУ, 2012. – Ч. 1. – С. 49-51.

7. Гультяева Т.А. Классификация сигналов, представляющих собой суммарный ток, потребляемый в электрической цепи, при наличии шума и близких параметров нагрузок / Т.А. Гультяева, Д.Ю. Коротенко // Информатика и проблема телекоммуникаций: материалы российской науч.-технич. конф. – Новосибирск: Изд-во СибГУТИ, 2012. – Т. I. – С. 23-26

8. Гультяева Т.А. Применение скрытых марковских моделей, kNN и SVM для задачи классификации режимов системы электроснабжения / Т.А. Гультяева, А.А. Попов, Д.Ю. Коротенко // Актуальные проблемы электронного приборостроения. АПЭП 2012: материалы XI междунар. науч.-технич. конф. – Новосибирск : Изд-во НГТУ, 2012. – том 6 – С. 36-41.

9. Гультяева Т.А. Исследование возможностей применения алгоритма k-ближайших соседей и метода опорных векторов при классификации последовательностей, порожденных скрытыми марковскими моделям / Т.А. Гультяева, Д.Ю. Коротенко // Сборник Научных трудов НГТУ. – Новосибирск : Изд-во НГТУ, 2011. – № 3 (65). – С. 45-55.

10. Гультяева Т.А. Исследование возможностей применения алгоритма k-ближайших соседей и метода опорных векторов для классификации сигналов, порожденных скрытыми марковскими моделями / Т.А. Гультяева, Д.Ю. Коротенко // Наука. Технологии. Инновации - 2011 : материалы Всерос. науч. конф. молодых ученых. – Новосибирск : Изд-во НГТУ, 2011. – Ч. 1. – С. 242–246.

11. Гультяева Т.А. Классификация последовательностей, подверженных действию помех с характеристиками, зависящими от скрытых состояний / Т.А. Гультяева, А.А. Попов // Сборник Научных трудов НГТУ. – Новосибирск : Изд-во НГТУ, 2011. – № 1 (63). – С. 59-68.

12. Гультяева Т.А. Классификация последовательностей, порожденных близкими скрытыми марковскими моделями, при наличии шума / Т.А. Гультяева, А.А. Попов // Технические науки: проблемы и перспективы : материалы междунар. заоч. науч. конф. – Санкт-Петербург : Изд-во Реноме, 2011. – № 1. – С. 37-41.

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

14. Гультяева Т.А. Классификация последовательностей, порождённых скрытыми марковскими моделями, при наличии шума / Т.А. Гультяева, А.А. Попов // Математические методы распознавания образов-15 : материалы всеросс. конф. – Москва: Изд-во МАКС Пресс, 2011. – С. 211–214.

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

16. Гультяева Т.А. Вычисление первых производных от логарифма функции правдоподобия для скрытых марковских моделей / Т.А. Гультяева // Сборник Научных трудов НГТУ. – Новосибирск : Изд-во НГТУ, 2010. – № 2 (60). – С. 39-46.

17. Гультяева Т.А. Особенности вычисления первых производных от логарифма функции правдоподобия для скрытых марковских моделей при длинных сигналах / Т.А. Гультяева // Сборник Научных трудов НГТУ. – Новосибирск : Изд-во НГТУ 2010. – № 2 (60). – С. 47-52.

18. Гультяева, Т.А. Повышение классификационных свойств скрытых марсковских моделей / Т.А. Гультяева // Информатика и проблема телекоммуникаций: материалы российской науч.-технич. конф. – Новосибирск: Изд-во СибГУТИ, 2010. – Т. I. – С.52-54.

Отпечатано в типографии
Новосибирского государственного технического университета
630073, г. Новосибирск, пр. К. Маркса, 20
Тел./факс (383) 346-08-57
Формат 60 х 84/16. Объем 1 п.л. Тираж 110 экз.
Заказ 1486. Подписано в печать 18.11.2013 г.