На правах рукописи

Кудинов Павел Юрьевич

АДАПТИВНЫЕ МЕТОДЫ ИЗВЛЕЧЕНИЯ ИНФОРМАЦИИ ИЗ СТАТИСТИЧЕСКИХ ТАБЛИЦ, ПРЕДСТАВЛЕННЫХ В ТЕКСТОВОМ ВИДЕ

Специальность:

05.13.17 – теоретические основы информатики

АВТОРЕФЕРАТ

диссертации на соискание учёной степени

кандидата технических наук

Москва – 2011

Работа выполнена в Учреждении Российской академии наук
Вычислительный центр им. А. А. Дородницына РАН.

Научный руководитель:	доктор физико-математических наук **Воронцов Константин Вячеславович**.
Официальные оппоненты:	доктор технических наук **Миркин Борис Григорьевич**, кандидат технических наук **Лещинер Дмитрий Роальдович**.
Ведущая организация:	Московский физико-технический институт (государственный университет).

Защита диссертации состоится «16» февраля 2012 г. в 14 часов на заседании диссертационного совета Д 002.017.02 при Учреждении Российской академии наук Вычислительный центр им. А. А. Дородницына РАН по адресу: 119333, г. Москва, ул. Вавилова, д. 40.

С диссертацией можно ознакомиться в библиотеке ВЦ РАН.

Автореферат разослан «29» декабря 2011 г.

Учёный секретарь диссертационного совета
Д 002.017.02, д.ф.-м.н., профессор *В. В. Рязанов*

Общая характеристика работы

Диссертационная работа посвящена проблеме извлечения статистических показателей из таблиц, представленных в текстовом виде. Предлагается методология полуавтоматической обработки исходных данных, основанная на динамическом обучении с привлечением эксперта. Исследуются инкрементные алгоритмы машинного обучения, не совершающие ошибок на обучающей выборке. Описывается полуавтоматическая система извлечения статистических показателей из таблиц, разработанная на основе предложенной методологии.

Актуальность темы. Социальная, экономическая, демографическая, финансовая статистика собирается и публикуется в бумажном и электронном виде различными организациями. Многие источники, например Росстат, ВЦИОМ, OECD, банки и финансовые организации предоставляют статистические данные в табличном виде (Рис. 1). Число таблиц, созданных ежегодно, измеряется сотнями тысяч. При этом в разных источниках могут использоваться разные термины и структуры таблиц для описания одних и тех же явлений. Это осложняет поиск, агрегацию и анализ динамики изменения статистических показателей.

В настоящее время не существует единого удобного способа поиска статистической информации по всем источникам. Организациям, которые занимаются анализом статистики, приходится обрабатывать таблицы вручную или с помощью примитивных программных средств, адаптация которых к быстро меняющимся источникам требует большого количества ресурсов. При этом обработка каждой новой коллекции таблиц занимает много времени, что ограничивает возможность оперативного анализа статистических данных. Таким образом, создание поисковой системы над множеством доступных источников является актуальной проблемой.

Исходными данными для поисковой системы являются коллекции таб-

лиц, получаемые из интернета с помощью роботов (crawler), либо загружаемые вручную. Поиск производится не по документам, как в классических поисковых системах, а по *статистическим показателям*, записанным в статистических таблицах. Каждый показатель характеризуется названием измеряемого статистического явления, единицей измерения, периодом времени и значением. Основной функцией системы поиска статистической информации является выдача, агрегирование, фильтрация, группировка статистических показателей и представление их в виде графиков, диаграмм или таблиц. В качестве запросов может использоваться один или несколько фрагментов названий статистических показателей, период времени, регионы и т. п.

Распределение численности занятых в экономике регионов Российской Федерации по возрастным группам в 2000 г. (в процентах)

	Всего	в том числе в возрасте, лет						Средний возраст, лет
		до 20	20-29	30-39	40-49	50-59	60-72	
Российская Федерация	**100**	**2,0**	**21,5**	**27,2**	**30,3**	**14,1**	**5,0**	**39,3**
Центральный федеральный округ	**100**	**1,6**	**20,0**	**26,6**	**30,1**	**15,7**	**6,1**	**40,1**
Белгородская область	100	1,8	19,8	28,4	30,5	12,6	6,9	39,9
Брянская область	100	1,9	22,8	27,7	30,7	12,3	4,6	38,8
Владимирская область	100	2,2	21,0	26,5	30,6	14,4	5,2	39,4
Воронежская								

Рис. 1. Пример статистической таблицы.

Известные методы информационного поиска ориентированы на обработку текстовых документов или изображений и не предназначены для обработки информации, представленной в табличной форме. Для создания поисковой системы по статистическим данным необходимо сначала решить задачу извлечения статистических показателей из произвольных таблиц. Основной проблемой является многообразие возможных способов записи одних и тех же показателей в таблицах. Исходные таблицы могут содержать неполные, противоречивые, по-разному агрегированные данные, строиться на разной терминологии и иметь разнородную, плохо формализованную структуру. Су-

щественной проблемой является большое количество опечаток и ошибок, которые появляются на этапе набора таблиц человеком или при применении автоматических методов распознавания текстов (OCR — Optical Character Recognition). Такие особенности исходных данных значительно снижают релевантность поисковой выдачи и делают задачи верификации и агрегирования статистических показателей трудновыполнимыми.

В настоящее время много научных исследований посвящено извлечению данных из таблиц. Но большинство методов либо не являются обучаемыми, либо ориентированы на узкое множество исходных таблиц. Адаптация этих методов к более широкому классу таблиц приведёт к неуправляемому росту сложности программной реализации. Методы, основанные на обучении по фиксированной выборке, неэффективно используют время эксперта, т.к. требуют разметки представительной обучающей выборки, необходимый объём которой трудно оценить заранее.

Более рациональным представляется режим диалога, при котором система предъявляет размеченные ею таблицы, а эксперт только исправляет допущенные системой ошибки. Эти исправления формируют обучающую выборку в режиме адаптивного или инкрементного обучения (incremental learning). При этом к алгоритму обучения предъявляются *требование корректности* — он не должен совершать ошибок на всех исправлениях, сделанных экспертом ранее.

Корректные инкрементные методы обучения, способные эффективно обучаться по выборкам длиной в сотни тысяч объектов и более, в литературе практически не изучались.

Таким образом, актуальной задачей является как создание автоматизированной системы для извлечения статистических показателей из текстовых таблиц, так и разработка корректных инкрементных методов обучения, достаточно эффективных на выборках большого объёма.

Цель диссертационной работы состоит в создании общей методологии автоматизированной обработки информации с привлечением экспертов и разработке корректных инкрементных методов классификации.

Методы исследования. Извлечение информации из статистических таблиц представляется в виде последовательности задач распознавания: определения типа ячеек, особенностей структуры таблиц, выделения названия таблицы и содержимого каждой ячейки. Некоторые из них формулируются как задачи классификации, для решения которых в данной работе применяются инкрементные методы машинного обучения и методы оптимизации для улучшения качества классификации. Для распознавания содержимого ячеек таблицы используются методы информационного поиска и анализа текстов.

На защиту выносятся следующие научные результаты и положения:

1. Методология построения обучаемых систем извлечения информации из плохо структурированных текстовых данных, основанная на применении корректных инкрементных алгоритмов классификации.

2. Алгоритм инкрементного обучения композиций случайных деревьев (Random Incremental Forest, RIF).

3. Методы и технология извлечения статистических показателей из таблиц, представленных в текстовом виде.

Научная новизна. Постановка задачи извлечения статистических показателей из больших коллекций разнородных таблиц, представленных в текстовом виде, является новой. Автором предложена методология разметки обучающей выборки и корректный инкрементный алгоритм машинного обучения, основанный на решающих деревьях и их композициях.

Научная значимость. В работе предложены и исследованы корректные алгоритмы инкрементного обучения, применимые для решения широкого класса задач.

Практическая значимость. Предложенная технология является основой для построения поисковой системы по статистическим таблицам. Использование инкрементных методов позволяет существенно сократить трудозатраты по формированию обучающих выборок.

Модули извлечения данных из таблиц были использованы в проектах Международного научно-образовательного Форсайт-центра Института статистических исследований и экономики знаний Национального исследовательского университета «Высшая школа экономики» при обработке экспертных оценок, представленных в виде статистических таблиц различного типа.

Апробация работы. Результаты работы докладывались и обсуждались на следующих научных конференциях:

- XI Всероссийская объединенная конференция «Интернет и современное общество» IMS–2008 (Санкт-Петербург, 2008 год);
- XVI международная научная конференция студентов, аспирантов и молодых учёных «Ломоносов–2009» (Москва, 2009 год);
- 14-я всероссийская конференция «Математические методы распознавания образов» (Суздаль, 2009 год);
- 8-я международная конференция «Интеллектуализация обработки информации» ИОИ'10 (Пафос, Республика Кипр, 2010 год);
- 6-я международная конференция «Служба экономических и социологических данных» ESDS-2010 (Лондон, Великобритания, 2010 год);
- 15-я всероссийская конференция «Математические методы распознавания образов» (Петрозаводск, 2011 год).

Описания отдельных результатов работы включены в отчёты по проектам РФФИ №№08-07-00305-а, 10-07-00609-а, 11-07-00480-а.

Личный вклад. Вклад автора работы в результаты, выносимые на защиту, является определяющим.

Публикации. Материалы диссертации опубликованы в 7 статьях [1–7], из них 2 работы [6, 7] — в журналах, включённых в Перечень ведущих рецензируемых научных журналов и изданий.

Структура и объём диссертации. Работа состоит из оглавления, введения, трёх глав, заключения и списка литературы. Содержание работы изложено на 105 страницах. Список литературы включает 46 наименований. Текст работы иллюстрируется 39 рисунками и 7 таблицами.

Содержание работы

Во введении обоснована актуальность диссертационной работы, сформулированы цели и задачи, аргументирована научная значимость исследования, представлены результаты и положения, выносимые на защиту, приведена краткая структура диссертации.

В первой главе описывается общая проблематика извлечения статистической информации из таблиц, обсуждаются известные подходы. Приводится концепция построения поисковой системы по статистическим таблицам. Она состоит из нескольких подсистем, реализация каждой из которых является сложной научно-технической задачей. Критически важной частью поисковой системы является система извлечения статистических показателей из таблиц, представленных в текстовом виде.

В разделе 1.1 описываются исходные таблицы и основные принципы построения системы поиска по статистической информации.

Система поиска статистической информации состоит из нескольких основных подсистем:

- подсистема поиска новых источников данных (статистических таблиц);
- подсистема извлечения статистических показателей из найденных таблиц;
- подсистема поиска релевантных статистических показателей и их отображения.

Решение общей задачи извлечения статистических показателей из таблиц сводится к последовательному решению следующих задач:

1) распознавание типа ячеек;
2) распознавание суперстрок;
3) распознавание вложенных ячеек;
4) извлечение названия таблицы из окружающего текста;
5) поиск содержимого ячеек и названия таблицы в словарях;
6) извлечение периода времени;
7) извлечение единиц измерения;
8) построение статистического показателя.

Задачи 1), 2), 3) являются задачами классификации, и для их решения предлагается использовать методы обучения по прецедентам. Объектами в этих задачах являются соответственно: ячейки, строки и некоторые пары соседних ячеек таблиц. Заметим, что если число таблиц (при самом скромном подсчёте) имеет порядок 10^4, то суммарное число ячеек — уже порядка 10^6. Таким образом, к эффективности применяемых методов обучения предъявляются серьёзные требования.

Основной проблемой является определение логической структуры статистической таблицы. Бóльшая часть таблиц содержит блоки ячеек описания сверху и слева. Такие таблицы относятся к классу таблиц *простой структуры* (рис. 2), т. к. для каждой ячейки данных множество ячеек описания

определяется тривиальным образом — необходимо взять все ячейки описания из данной строки и столбца.

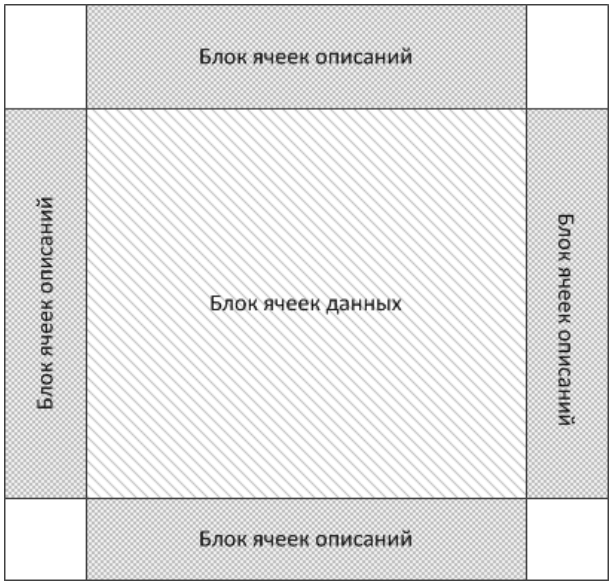

Рис. 2. Статистическая таблица простой структуры.

Таблица может быть разделена на несколько частей в соответствии со значениями некоторого показателя. Например, данные по мужской и женской занятости, данные за разные годы, абсолютные и относительные данные одних и тех же показателей. Для совмещения таких данных в одной таблице составители часто используют суперстроки (рис. 3).

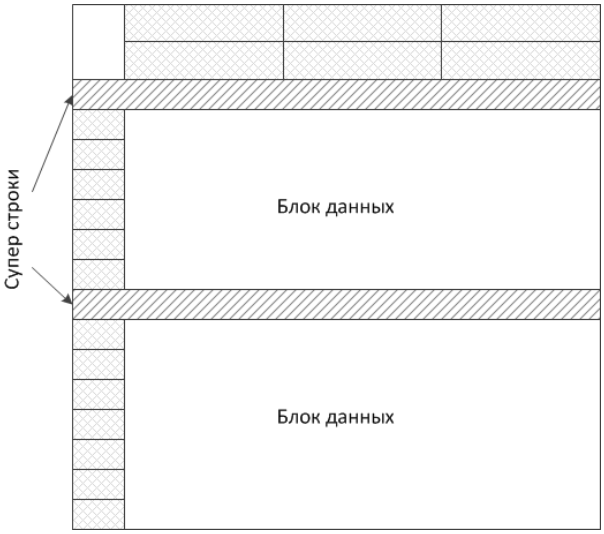

Рис. 3. Статистическая таблица с суперстроками.

Ещё одним распространённым приёмом оформления таблиц является использование вложенных ячеек, когда несколько ячеек сдвигаются на один уровень вправо (рис. 4). Этот приём часто используется в таблицах Росстата. Его игнорирование приводит к неправильному пониманию сути отражённой в таблице информации. Таблицы с несколькими уровнями вложенности встречаются редко и в настоящей работе не рассматриваются.

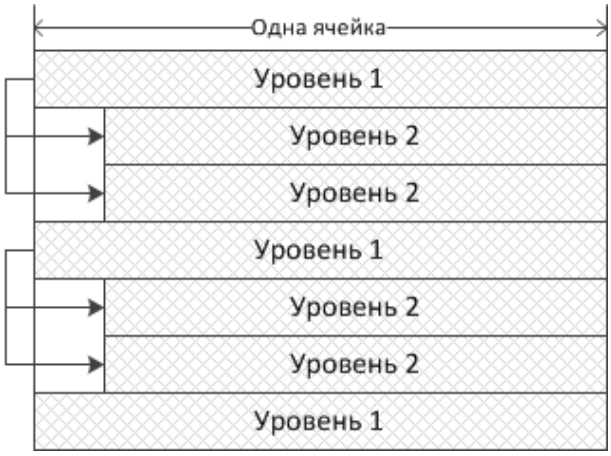

Рис. 4. Фрагмент таблицы с вложенными ячейками.

Даты в статистических таблицах представляются в числовой и текстовой формах, а также могут быть подвергнуты уточнениям различных типов, выраженным в текстовой форме. Например, статистические данные по учебным заведениям обычно заданы для учебного года. Задача состоит в том, чтобы для каждого статистического показателя получить интервал времени $[t_1, t_2]$.

В тексте одной ячейки может содержаться несколько *ключей* (элементарных характеристик статистических явлений). Поэтому задача построения названия извлекаемого статистического показателя формулируется как задача поиска всевозможных словесных представлений ключей в текстах ячеек описания, определяющих этот показатель.

В разделе 1.2 предлагается концепция полуавтоматической системы извлечения статистических показателей из таблиц, основанной на динамическом обучении.

При динамическом обучении объекты появляются по одному, причём каждый объект обрабатывается только один раз. Динамическое обучение состоит из последовательности чередующихся шагов классификации и дообучения. На стадии классификации очередного объекта алгоритму не известно значение целевого признака. После того как эксперт произвёл верификацию результатов классификации, алгоритму передаётся правильный ответ, который используется для модификации параметров алгоритма (дообучения).

Эксперт исправляет только ошибки классификации, что существенно уменьшает объём работы по сравнению с обычной практикой, когда обучающая выборка формируется заранее (offline learning). Алгоритм классификации, используемый для решения этой задачи, должен удовлетворять *требованию корректности*, то есть при последующих модификациях он не должен делать ошибок на ранее классифицированных объектах.

Во второй главе вводятся основные понятия, определения и обозначения, описываются особенности статистических таблиц, приводится постановка задачи извлечения статистических показателей, излагаются все этапы её решения. Для этапов, представляющих из себя задачи распознавания, приводятся постановки соответствующих задач, определяются признаки классифицируемых объектов. Описанные задачи распознавания предлагается решать инкрементными алгоритмами машинного обучения. Проводится сравнение инкрементных алгоритмов и обосновывается выбор алгоритма построения дерева решений ITI. Предлагается новый корректный алгоритм построения композиции случайных деревьев решений (RIF).

В разделе 2.1 вводятся основные понятия и определения.

Ключом k будем называть элементарную, неделимую характеристику измеряемого явления. Множество всех ключей обозначим через $\mathcal{K}$. Каждому ключу $k_i \in \mathcal{K}$ соответствует множество его допустимых словесных описаний D_i. Оно включает все словесные эквиваленты одной и той же статистиче-

ской характеристики, представленной в текстовом виде. Множества словесных описаний могут быть получены из Общероссийских классификаторов или наполнены вручную.

Статистическим показателем будем называть четвёрку $s = \langle K, U, d, v \rangle$, определяющую множество ключей $K = \{k_1, \ldots, k_n\} \subset \mathcal{K}$, единицу измерения $U \subset K$, интервал времени $d = [t_1, t_2]$, где t_1, t_2 — даты, и значение $v \in \mathbb{R}$.

Множество K полностью описывает смысл (географическое положение, вид экономической деятельности, вид продукции и т.п.) измеряемого социально-экономического явления. Пространство всех статистических показателей будем обозначать через $\mathcal{I}$.

Рассмотрим квадратную сетку $G^{M \times N}$ из M строк и N столбцов и зададим её полное покрытие непересекающимися прямоугольными областями. Элемент покрытия будем называть *ячейкой*. Множество всех ячеек обозначим через C. Каждой ячейке $c \in C$ поставим в соответствие текстовую строку $text(c)$, возможно пустую, которую назовём *содержимым ячейки* или её *текстом*. Множество C задаёт *физическую структуру* таблицы.

Определение 1. *Таблицей будем называть пару* $\langle G^{M \times N}, C \rangle$.

Пространство всех таблиц обозначим через $\mathcal{T}$.

Будем полагать, что каждая ячейка c имеет один из трёх типов: ячейка данных, ячейка описания или неинформативная ячейка. Определим отображение $\mathfrak{R}\colon C_V \to 2^{C_K}$, где C_V — множество ячеек данных и C_K — множество ячеек описаний. Отображение $\mathfrak{R}$ ставит в соответствие каждой ячейке данных множество ячеек описания, т. е. задаёт *логическую структуру* таблицы.

Определение 2. *Статистической таблицей будем называть четвёрку*

$$\langle G^{M \times N}, C_V, C_K, \mathfrak{R} \rangle.$$

Множество всех статистических таблиц обозначим через $\mathcal{T}_s$. Будем полагать, что если таблица T является статистической, то для неё определено отображение $\mathfrak{S} : \mathcal{T} \to \mathcal{T}_s$.

В разделе 2.2 приводится постановка задачи извлечения статистических показателей из таблиц, которая состоит в том, чтобы построить отображение $\mathfrak{M} : \mathcal{T} \to 2^{\mathcal{I}}$.

Решение этой задачи разбивается на два основных этапа: распознавание статистической таблицы, т. е. построение отображения $\mathfrak{S}$, а затем построение отображения $\mathfrak{I} : \mathcal{T}_s \to 2^{\mathcal{I}}$.

Первый этап состоит из следующих шагов.

1. Распознавание типа каждой ячейки, т. е. построение множеств C_V и C_K.
2. Распознавание суперстрок.
3. Распознавание вложенных ячеек.
4. Чтение таблицы, т. е. построение отображения $\mathfrak{R}$.

Второй этап состоит из следующих шагов.

1. Чтение ячеек описания — извлечение ключей из текста всех ячеек описания.
2. Извлечение названия таблицы из текста.
3. Извлечение периода времени и единицы измерения.
4. Построение множества статистических показателей.

Распознавание типа ячеек. Положение ячейки $c \in C$ в таблице описывается координатами левого верхнего $\big(r_1(c), c_1(c)\big)$ и правого нижнего $\big(r_2(c), c_2(c)\big)$ углов прямоугольника по сетке $G^{M \times N}$. Для каждой ячейки $c \in C$ определим следующие признаки:

1) $f_1(c)$ — количество чисел в $text(c)$;

2) $f_2(c)$ — количество слов в $text(c)$;

3) $f_3(c)$ — количество символов в $text(c)$;

4) $f_4(c) = \big(r_1(c) + r_2(c)\big)/2M$ — вертикальное положение;

5) $f_5(c) = \big(c_1(c) + c_2(c)\big)/2N$ — горизонтальное положение ячейки c;

6) $f_6(c) = r_2(c) - r_1(c)$ — число строк сетки, занимаемых ячейкой c;

7) $f_7(c) = c_2(c) - c_1(c)$ — число столбцов, занимаемых ячейкой c.

Классификация суперстрок. Рассмотрим задачу классификации строк таблицы на два класса: «обычная строка» и «суперстрока». Для каждой строки $S \subset C$ будем строить следующий набор признаков:

1) $f_1(S)$ — количество ячеек в строке;

2) $f_2(S) = N$ — ширина таблицы;

3) $f_3(S)$ — высота строки;

4) $f_4(S)$ — количество пустых ячеек;

5) $f_5(S) = \big(\min\limits_{c \in S} c_1(c) + \max\limits_{c \in S} c_2(c)\big)/2N$ — количество непустых ячеек.

Вложенные ячейки. Для определения вложенности решается задача классификации, в которой объектами являются пары последовательно идущих ячеек $p = (x_1, x_2)$ в левом блоке ячеек описания, разделённых на три класса: «ячейки находятся на одном уровне», «x_2 сдвинута вправо относительно x_1» и «x_2 сдвинута влево относительно x_1». Для этих объектов вычисляется следующий набор признаков:

1) $f_1(p)$ — текст x_1 заканчивается на «:»;

2) $f_2(p)$ — количество начальных пробельных символов в тексте x_2;

3) $f_3(p)$ — тип первого непробельного символа в x_2: «цифра», «буква» или «знак»;

4) $f_4(p)$ — первая буква в x_1 является прописной;

5) $f_5(p)$ — первая буква в x_2 является прописной;

6) $f_6(p) = r_2(x_1) - r_1(x_1)$ — высота x_1;

7) $f_7(p) = c_2(x_2) - c_1(x_2)$ — ширина x_2.

Чтение ячейки описания. Задача состоит в том, чтобы каждой ячейке-описанию $c \in C_K$ поставить в соответствие множество ключей $K(c) \subset \mathcal{K}$, то есть построить отображение $\mathfrak{F} : C_K \to 2^{\mathcal{K}}$. При этом необходимо учитывать, что, во-первых, допустимы ошибки в словах, а во-вторых — разделение строки на ключи может быть не однозначно.

Текст ячейки описания будем рассматривать как последовательность слов $(x_1, \ldots, x_n)$. Для каждого слова x_i строится множество близких слов $\tilde{X}_i = \{x \in \mathcal{W} : \rho(x, x_i) < \tau\}$, $\mathcal{W}$ — все слова и словоформы, ρ — расстояние Левенштейна, τ — параметр. После этого строится семейство всех подстрок строки x, включающее всевозможные комбинации близких слов.

$$\mathcal{X} = \{(\tilde{x}_i, \ldots, \tilde{x}_j),\ 1 \leqslant i \leqslant j \leqslant n, \tilde{x}_l \in \tilde{X}_l, \forall l = \overline{i,j}\}.$$

Обозначим через $\theta(\tilde{x})$ множество номеров слов, входящих в строку $\tilde{x}$.

Построим $\mathcal{X}_\mathcal{D} = \mathcal{X} \cap \mathcal{D}$ — множество всех строк $\mathcal{X}$, являющихся описаниями ключей $\mathcal{D} = \bigcup\limits_{i=1\ldots|\mathcal{K}|} D_i$. Рассмотрим множество всех разбиений $\mathcal{X}$ на ключи:

$$\overline{\mathcal{X}_\mathcal{D}} = \{S \in 2^{\mathcal{X}_\mathcal{D}} : \forall s_1, s_2 \in S \Rightarrow \theta(s_1) \cap \theta(s_2) = \varnothing\}.$$

Обозначим множество номеров слов исходной строки x, не содержащихся ни в одной из строк множества $S \in \overline{\mathcal{X}_\mathcal{D}}$ через

$$\overline{\Theta}(S) = \{\overline{1,n}\} \setminus \bigcup_{s \in S} \theta(s), \alpha(s) = \sum_{x \in s} \rho(\tilde{x}, x).$$

Для оценки качества разбиения предлагается следующий функционал:

$$E(S) = \gamma|\overline{\Theta}(S)| + \sum_{s \in S} \alpha(s) \to \min_{S \in \overline{\mathcal{X}_\mathcal{D}}}.$$

Минимизация функционала осуществляется на основе перебора всех элементов множества $\overline{\mathcal{X}_\mathcal{D}}$.

Утверждение 2.1. *При полном переборе сложность построения множества $\mathcal{X}_\mathcal{D}$ равна $O(n^2 f^n |\mathcal{D}|)$, $f = \max\limits_{i=1,\ldots,n} |\tilde{X}_i|$.*

Для построения множества $\mathcal{X}_\mathcal{D}$ в работе предлагается использовать обобщённые суффиксные деревья, построенные над множеством $\mathcal{D}$ в алфавите номеров слов $\Sigma_\mathcal{W} = \{1, \ldots, |\mathcal{W}|\}$. Доказывается следующее утверждение:
Утверждение 2.2. *При использовании суффиксных деревьев сложность построения множества $\mathcal{X}_\mathcal{D}$ равна $O(n^2 f^n)$, $f = \max\limits_{i=1,\ldots,n} |\tilde{X}_i|$.*

В разделе 2.3 рассматриваются корректные инкрементные алгоритмы классификации, основанные на деревьях решений. Обозначим всё множество объектов через X, а множество признаков — через $\mathcal{F} = \{f_1, ..., f_M\}$, где $f_i : X \to D_{f_i}$. Если $|D_{f_i}| < \infty$, то f_j — номинальный признак, иначе — числовой. Множество классов обозначим через Y. Пусть дана обучающая выборка $X^l = \{(x_i, y_i)_{i=1}^l\}$, состоящая из l пар объект-ответ.

Определение 3. *Алгоритм называется корректным на выборке X^l, если классификация каждого объекта $x \in X^l$ безошибочна.*

Алгоритм Incremental Tree Induction, ITI (Utgoff, 1997) предполагает инкрементное построение бинарного дерева решений. В узловой вершине ν инкрементного дерева хранится пятёрка $\langle L_\nu, R_\nu, \beta_\nu, X_\nu, s_\nu \rangle$, где L_ν, R_ν, β_ν — левая, правая вершина и предикат соответственно, $X_\nu \subset X^l$ — список объектов, прошедших через узел ν, s_ν — состояние узла. Если через него прошли новые объекты, то $s_\nu = 1$, иначе $s_\nu = 0$. Будем рассматривать $\beta_\nu(x) = [f_j(x) < a]$ в случае числового признака и $\beta_\nu(x) = [f_j(x) = a]$ — в случае номинального. Каждому листу ν соответствует метка класса.

При добавлении нового объекта x он проходит по дереву от корня к листьям в соответствии с предикатами, установленными в узлах. Если он

попадает в лист ν с объектами другого класса, то ν превращается в узел. Условие для нового узла выбирается так, что одному из листьев будет соответствовать множество объектов X_ν, а второму — добавляемый объект x с соответствующими метками классов.

Лемма 2.1. *Решающее дерево, построенное в результате последовательного добавления объектов из непротиворечивой выборки, является корректным алгоритмом классификации.*

Дерево, построенное инкрементно, зависит от порядка добавления объектов и может существенно отличаться от оптимального. Для решения этой проблемы используется *операция транспозиции* дерева, заключающаяся в периодическом поиске лучшего предиката для каждого узла дерева. Временны́е затраты на транспозицию в среднем меньше, чем на полное перестроение дерева, но её выполнение затрудняет практическое применение алгоритма.

В разделе 2.4 описывается новый корректный инкрементный алгоритм классификации, основанный на построении композиции деревьев.

По начальной обучающей выборке строится композиция решающих деревьев $RIF = \{\langle RITree_i, M_i\rangle\}_{i=1}^{p}$, где $RITree_i$ — инкрементное дерево, построенное по некоторому набору признаков $M_i \subseteq \mathcal{F}$. Этот набор может быть получен случайным образом или в результате отбора признаков. Для построения каждого дерева используется одна и та же исходная выборка. Результатом классификации композиции является результат голосования входящих в неё деревьев.

Для каждого i-го дерева генерируется поднабор M_i из $M \leqslant |\mathcal{F}|$ признаков. Каждое дерево строится на основе алгоритма ITI с небольшим изменением — поиск наилучшего условия ветвления в узлах дерева осуществляется не по всему подможеству признаков, а по случайному признаку из M_i.

Лемма 2.2. *Композиция случайных инкрементных деревьев RIF, построенная по непротиворечивой выборке, является корректным алгорит-*

мом классификации.

Получающиеся в ходе инкрементного обучения деревья могут существенно отличаться по качеству. В случае длинных обучающих выборок можно осуществить отбор деревьев по критерию качества классификации новых (контрольных объектов).

Допустим, ошибка дерева является случайной величиной из распределения Бернулли с параметром p. Тогда верхняя доверительная граница для p с уровнем доверия α, согласно теореме Муавра–Лапласа, равна

$$\bar{p} = \frac{m}{n} + \Phi_\alpha \sqrt{\frac{m(n-m)}{n^3}},$$

где Φ_α — α-квантиль стандартного нормального распределения, n — число объектов, классифицированных деревом после последнего перестроения, m — число ошибок на контрольных объектах. Полученная оценка используется для отбора деревьев при некотором фиксированном α.

Процесс отбора деревьев представляет собой последовательность чередующихся операций удаления нескольких худших деревьев и добавления новых. Множества признаков M_i новых деревьев выбирается на основе признаков лучших деревьев. Построение новых деревьев происходит по всей доступной на данный момент выборке. Доказывается теорема о корректности полученной в результате отбора деревьев композиции.

Теорема 2.1. *Композиция деревьев, полученная в результате работы процедуры отбора деревьев и построенная по непротиворечивой выборке, является корректным алгоритмом классификации.*

В третьей главе описывается архитектура программной системы извлечения статистических показателей, особенности реализации процедуры инкрементного обучения деревьев и их композиций и интерфейса пользователя. Приводятся результаты экспериментов.

В разделе 3.1 приводится архитектура приложения, реализующего про-

цедуру динамического извлечения статистической информации из таблиц.

В разделе 3.2 приводятся результаты экспериментов. Эксперименты проводились для сравнения нового алгоритма RIF с алгоритмом ITI на задачах репозитория UCI и реальной выборке из 600 таблиц Росстата из коллекций «Регионы России, 2008» и «Финансы России, 2010».

Для оценки качества работы алгоритма использовался скользящий контроль. Инкрементное обучение запускалось несколько раз на одной выборке с её случайным перемешиванием. Сначала алгоритм обучался на небольшой начальной части обучающей выборки. Затем все объекты классифицировались по очереди. После этого объект подавался на дообучение. Строились кривые обучения — зависимости средней частоты ошибок от длины обучающей выборки.

В результате экспериментов оказалось, что новый алгоритм построения случайного инкрементного леса с отбором деревьев работает лучше на 7 задачах из 8 по сравнению с алгоритмом ITI с транспозицией и без. Доля ошибок на задачах классификации при извлечении статистических показателей для алгоритма RIF составила $0,07\%$ на задаче распознавания типа ячеек, $0,03\%$ на задаче распознавания суперстрок и $2,8\%$ для задачи распознавания вложенных ячеек.

В заключении сформулированы основные результаты диссертационной работы.

Список публикаций

[1] Интегрированная база данных по социально-демографической статистике: ресурс и пользовательские сервисы для поддержки гуманитарных исследований / А. В. Богомолова, О. И. Карасев, П. Ю. Кудинов и др. // Труды XI Всероссийской объединенной конференции «Интернет и современное общество» IMS-–2008 (Санкт-Петербург, 28–30 октября 2008 года). — СПб.: Факультет филологии и искусств СПбГУ, 2008. — С. 58–59.

[2] *Кудинов П. Ю.* Задача распознавания статистических таблиц // Доклады 14-й Всероссийской конференции «Математические методы распознавания образов» ММРО-2009. — М.: МАКС Пресс, 2009. — С. 552–555.

[3] *Кудинов П. Ю.* Об одном подходе к организации системы распознавания таблиц, содержащих статистические данные // Материалы XVI Международной конференции студентов, аспирантов и молодых учёных «Ломоносов–2009». — М.: Издательский отдел факультета ВМиК МГУ; МАКС Пресс, 2009. — С. 44.

[4] *Кудинов П. Ю., Полежаев В. А.* Динамическое обучение распознаванию статистических таблиц // Доклады 8-й Международной конференции «Интеллектуализация обработки информации» ИОИ–2010 (Республика Кипр, г. Пафос, 17–24 октября 2010). — М.: МАКС Пресс, 2010. — С. 512–515.

[5] *Кудинов П. Ю., Полежаев В. А.* Инкрементное обучение деревьев решений в задаче распознавания структуры статистических таблиц // Доклады 15-й Всероссийской конференции «Математические методы распознавания образов» ММРО–2011. — М.: МАКС Пресс, 2011. — С. 593–596.

[6] *Кудинов П. Ю., Полежаев В. А.* Композиция случайных инкрементных

деревьев и восстановление структуры таблиц // ***Бизнес-информатика***. — 2011. — № 4(18). — С. 39–46.

[7] *Kudinov P. Y.* Extracting statistics indicators from tables of basic structure // ***Pattern Recognition and Image Analysis***. — 2011. — Vol. 21, no. 4. — Pp. 630–636.