САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ
УНИВЕРСИТЕТ

На правах рукописи

ШАЛЫМОВ Дмитрий Сергеевич

МАТЕМАТИЧЕСКОЕ ОБЕСПЕЧЕНИЕ ДЛЯ РАЗРАБОТКИ И АНАЛИЗА СИСТЕМ РАСПОЗНАВАНИЯ ОБРАЗОВ, ИСПОЛЬЗУЮЩИХ РАНДОМИЗИРОВАННЫЕ АЛГОРИТМЫ

05.13.11 — Математическое и программное обеспечение
вычислительных машин, комплексов и компьютерных сетей

АВТОРЕФЕРАТ
диссертации на соискание ученой степени
кандидата физико-математических наук

Санкт-Петербург
2009

Работа выполнена на Кафедре системного программирования Математико-механического факультета Санкт-Петербургского государственного университета.

Научный руководитель:	доктор физико-математических наук, профессор ГРАНИЧИН Олег Николаевич
Официальные оппоненты:	доктор технических наук, профессор ТИМОФЕЕВ Адиль Васильевич (Санкт-Петербургский институт информатики и автоматизации РАН)
	кандидат физико-математических наук, доцент КОСОВСКАЯ Татьяна Матвеевна (Санкт-Петербургский государственный университет)
Ведущая организация:	Институт проблем управления им. В.А.Трапезникова РАН (Москва)

Защита состоится "____" _____ 2009 года в ____ часов на заседании совета Д 212.232.51 по защите докторских и кандидатских диссертаций при Санкт-Петербургском государственном университете по адресу 198504, Санкт-Петербург, Петродворец, Университетский пр.,28, Математико-механический факультет.

С диссертацией можно ознакомиться в Научной библиотеке им.М.Горького Санкт-Петербургского государственного университета по адресу: 199034, Санкт-Петербург, Университетская наб. 7/9.

Автореферат разослан "____" _____ 2009 г.

Ученый секретарь
диссертационного совета Даугавет И. К.

Общая характеристика работы

Актуальность темы. На протяжении последних десятилетий в связи со стремительным развитием цифровых технологий наблюдается значительный рост объемов хранимых и перерабатываемых данных. Однако увеличение количества информации не означает непосредственного увеличения объемов знаний. В такой ситуации все более востребованными становятся новые математические методы, которые позволяли бы распознавать образы, структурировать информацию и находить объективные закономерности в больших объемах данных. Среди них важную роль при распознавании образов играют методы выявления классов (кластеров), способные работать в режиме реального времени. О популярности этих методов сегодня свидетельствует тот факт, что результат поиска по запросу термина "classification problem" в поисковой системе Google (на сентябрь 2009 года) составил более сорока трех миллионов страниц.

Современные алгоритмы теории распознавания образов, классификации и кластерного анализа базируются на работах С.А.Айвазяна, А.Я.Червоненкиса, В.Н.Вапника, Ф.Розенблатта, Р.А.Фишера, В.Н.Фомина, И.Форджи, К.Фукунаги, Дж.Хартигана, Дж.Хопфилда, Я.З.Цыпкина и др. Многие современные системы распознавания образов основаны на принципах нейронных сетей (С.Хайкин, Ф.Уоссерман, А.В.Тимофеев и др.)

Работоспособность различных алгоритмов разбиения множества данных на классы существенно зависит от количества классов (кластеров) и выбора первоначального разбиения. При априори неизвестном количестве кластеров В.Кржановским и И.Лаем, Дж.Дуном, Л.Хьюбертом и Дж.Шульцом, Р.Калинским и Дж.Харабазом, Е.Левине и Е.Домани, А.Бен-Гуром и И.Гийоном, А.Елизивом, В.Волковичем и др., Р.Тибширани и Г.Вальтером и др. активно разрабатываются методы устойчивой кластеризации, достаточно точно оценивающие количество кластеров в разнообразных прикладных задачах.

Общим недостатком традиционно используемых алгоритмов кластеризации является значительный рост вычислительной сложности при увеличении мощности исследуемого множества. В условиях многомерных задач и нарастающих объемов данных в современных работах М.Вадьясагара, Дж.Галафиори и М.Кампи, О.Н.Граничина, Ю.М.Ермольева, В.Я.Катковника, А.И.Кибзуна и Ю.С.Кана, Г.Кушнера и Г.Ина, Б.Т.Поляка, П.С.Щербакова и А.Б.Цыбакова, Дж.Спала и др. эф-

фективно используются новые рандомизированные алгоритмы, развивающие идеи методов случайного поиска и моделирования по методу Монте-Карло, детально исследованные в русскоязычной литературе С.М.Ермаковым, А.А.Жиглявским и В.Б.Меласом, А.Жилинскасом, Л.А.Растригиным и многими другими. Сложность целого ряда новых рандомизированных алгоритмов, в англоязычной литературе получивших название SPSA (Simultaneous Perturbation Stochastic Approximation), не существенно возрастает при росте размерности данных и, кроме того, они остаются работоспособными в условиях значительных неконтролируемых воздействий, которые трудно исключить в системах реального времени.

Наряду с развитием методов распознавания образов активно разрабатываются соответствующие средства программного обеспечения как для настольных и супер компьютеров, так и для встроенных систем. Наборы библиотек с алгоритмами кластеризации входят в Matlab, SPSS, Statistica, SAS Enterprise Miner и многие другие популярные пакеты прикладных программ. Сформировано несколько больших хранилищ данных (UCI Machine Learning Repository, GEMLeR, StatLib, KDD cups и др.) для тестирования работоспособности алгоритмов и решения практически важных задач. Вместе с тем для разработки и анализа новых пользовательских систем распознавания образов не создано удобного общедоступного средства.

Цель работы. Создание математического обеспечения для разработки и анализа систем распознавания образов, использующих рандомизированные алгоритмы, работоспособных в условиях большой размерности и при незначительных ограничениях на неконтролируемые возмущения.

Цель достигается в диссертации через решение следующих задач:

- разработать и обосновать для распознавания образов слов в речи прототип дикторонезависимй системы, основанной на использовании рандомизированного алгоритма стохастической аппроксимации типа SPSA;
- разработать и обосновать новый рандомизированный метод устойчивой кластеризации, работоспособный в режиме реального времени;
- создать программный комплекс для разработки и анализа систем распознавания образов, использующих рандомизированные алгоритмы.

Методы исследования. В диссертации применяются методы теории оценивания и оптимизации, функционального анализа, теории вероятностей и математической

статистики, имитационного моделирования и системного программирования.

Основные результаты. В работе получены следующие основные научные результаты:

1. На основе рандомизированного алгоритма стохастической аппроксимации (РАСА) и метода кепстральных коэффициентов тоновой частоты разработано программное средство для распознавания образов слов в речи. Исследованы свойства помехоустойчивости РАСА в задаче распознавания и установлены условия состоятельности доставляемых алгоритмом РАСА оценок.

2. Предложен новый рандомизированный метод определения количества кластеров в множесте данных, работоспособный в режиме реального времени.

3. Получены и теоретически обоснованы условия достоверности предложенного нового рандомизированного метода определения количества кластеров в множесте данных.

4. Создан новый программный комплекс для разработки и анализа систем распознавания образов, базирующихся на использовании рандомизированных алгоритмов классификации и кластеризации, обеспечивающий технологичность разработки новых систем распознавания образов. Проведена апробация предложенных в диссертации алгоритмов на данных различной природы.

Научная новизна. Все основные научные результаты диссертации являются новыми.

Теоретическая ценность и практическая значимость. Теоретическая ценность работы состоит в обогащении теории распознавания образов современными новыми знаниями о возможностях применения новых рандомизированных алгоритмов в задачах распознавания образов в условиях многомерности фазового пространства и наличия неконтролируемых нерегулярных возмущений.

Предложенные новые методы могут быть эффективно использованы в современных практических задачах. Созданный программный комплекс для разработки и анализа систем распознавания образов позволяет исследовать работоспособность новых методов классификации и кластеризации, а также анализировать пользовательские данные с помощью большого набора алгоритмов, подбирая для них наиболее подходящие параметры. Реализованные в ходе диссертационного исследования

приложения рандомизированных алгоритмов в задачах кластеризации данных и распознавания отдельных слов речи представляют собой самостоятельную практическую ценность.

Апробация работы. Материалы диссертации докладывались на внутренних семинарах кафедры системного программирования математико-механического факультета СПбГУ, на российских и международных конференциях по оптимизации, информатике и теории управления: The 3rd Int. IEEE Scientific Conf. on Physics and Control "PhysCon – 2007" (Potsdam, Germany, September 3-7, 2007), 5-я межд. научно-практическая конф. "Исследование, разработка и применение высоких технологий в промышленности" (Санкт-Петербург, Россия, 28-30 апреля, 2008), The 20th Int. Conf. "Continuous Optimization and Knowledge-Based Technologies (EUROPT-08)" (Neringa, Lithuania, May 20-24, 2008), Yalta Conf. on Discrete and Global Optimization (Yalta, Ukraine, August 1-3, 2008), The ERANIS Int. event in the fields of KBBE, ICT, NMP and Energy (Warsaw, Poland, October 10, 2008), III Межд. научно-практическая конф. "Современные информационные технологии и ИТ-образование" (Москва, Россия, 6-9 декабря, 2008), The 8th Int. Conf. on System Identification and Control Problems "SICPRO'09" (Moscow, Russia, January 26-30, 2009), XI конф. молодых ученых "Навигация и управление движением" (Санкт-Петербург, Россия, 10-12 марта, 2009), VI Всероссийская межвузовская конф. молодых ученых (Санкт-Петербург, Россия, 14-17 апреля, 2009), Spring Young Researchers' Colloquium on Software Engineering "SYRCoSE" (Moscow, Russia, May 28-29, 2009), Первая традиционная всероссийская молодежная летняя школа "Управление, информация и оптимизация" (Переславль-Залесский, Россия, 21-28 июня, 2009), VI школа-семинар молодых ученых "Управление большими системами" (Ижевск, Россия, 31 августа - 5 сентября, 2009).

По материалам диссертации было получено свидетельство об официальной регистрации программы для ЭВМ N 2007611711 "Программная система для обучения, перевода, распознавания арабского текста" от 23 апреля 2007 года. Результаты диссертации были частично использованы в работе по гранту РФФИ 09-04-00789-а. Доклад "Рандомизированные алгоритмы устойчивой кластеризации для динамически изменяющихся данных" на VI Всероссийской межвузовской конференции молодых ученых в СПбГУ ИТМО был отмечен дипломом "За лучший доклад аспиранта на секции". Проект "ИнтАн: Программный комплекс интеллектуального анализа данных", использующий во многом материалы диссертации, принял уча-

стие в смене “Инновации и Техническое творчество” в рамках молодежного форума Селигер-2009. Результаты диссертационной работы были представлены в проекте “Разработка программного комплекса кластерного анализа данных большого объема”, который победил в конкурсе “У.М.Н.И.К.” в 2009 году.

Публикации. Основные результаты диссертации опубликованы в работах [1–18]. Из них три публикации [12,17,18] в журналах из перечня ВАК. Работы [2,3,8,9,12,14] написаны в соавторстве. В работах [2,3,8,9,12] О.Н.Граничину принадлежат общие постановки задач, а Д.С.Шалымову – реализация описываемых методов, создание демонстрационных примеров и программных средств. В [14] Д.С.Шалымов является автором I–VIII секций, К.Скрыгану принадлежит участие в реализации вычислительного ядра и соавторство в IV секции, посвященной организации его внутренней структуры, Д.Любимову принадлежит участие в создании демонстрационных примеров, проиллюстрированных на рис. 3–5.

Структура и объем диссертации. Диссертация состоит из введения, трех глав, заключения, списка литературы, включающего 136 источников. Текст занимает 126 страниц, содержит 34 рисунка и две таблицы.

Содержание работы

Во **введении** обосновывается актуальность тематики диссертационной работы и кратко излагаются ее основные результаты.

В **первой главе** “Задачи распознавания образов, классификации и кластеризации” вводятся основные понятия и формальные постановки задач исследований предметной области, дается исторический контекст развития итеративных алгоритмов кластеризации, рассматриваются примеры применения классификации и кластеризации в задачах распознавания образов.

В п. 1.1 анализируются типичные задачи распознавания образов, которые достаточно часто трактуются как классификация входных сигналов (стимулов, объектов). В процессе классификации обнаруживаются признаки, которые характеризуют группы объектов исследуемого набора данных – классы (кластеры). По этим признакам каждый сигнал можно отнести к тому или иному классу. Результатом кластеризации является разбиение сигналов на группы в условиях, когда классы заранее не определены. Получающееся разбиение естественным образом характе-

ризует структуру множества данных и может быть использовано в дальнейшем для ее определения.

В п. 1.2 описывается формальная постановка задачи. С содержательной точки зрения процесс распознавания образов трактуется как классификация входных сигналов x из некоторого множества X, заключающаяся в построении правила сопоставления каждой точке $x \in X$ некоторого образа (класса) X_k. (В диссертации для упрощения рассматриваются только задачи однозначной классификации, хотя это ограничение и не носит принципиального характера и может быть расширено.) Выбор правила классификации порождает разбиение множества X на классы. Будем считать, что правило классификации однозначно определяется конечномерным набором η и задана функция $l(\eta)$ возвращающая количество классов при классификации по правилу η. Всякий способ классификации η связан с потерями, которые обычно характеризуются с помощью штрафных функций стоимости $q^k(x, \eta)$ при отнесении точки x к классу с номером k. В типичных случаях, когда X — вещественное векторное пространство, значения штрафных функций $q^k(x, \eta)$ возрастают при удалении x от центра соответствующего образа (класса).

В диссертации рассматривается следующее правило классификации: при заданных η и функциях $q^k(x, \eta)$, $k = 1, \ldots, l(\eta)$, входной сигнал x из множества X относится к тому классу X_k с наименьшим номером $k = k(x)$, для которого значение соответствующей штрафной функции $q^k(x, \eta)$ минимально

$$k(x) = \min\{\arg \min_{i \in 1:l(\eta)} q^i(x, \eta)\}.$$

Например, если на X задана норма $\|\cdot\|$ и в множестве данных l классов, а η^l — набор векторов центров классов: $\eta^l = (\theta^1, \theta^2, ..., \theta^l)$, то можно считать $\eta = \{l, \eta^l\}$ и в качестве $q^k(x, \eta)$ можно задать расстояние до центра k-го класса θ^k: $q^k(x, \eta) = \left\|x - \theta^k\right\|^2$. При этом множество X разбивается на l классов $X_1(\eta), X_2(\eta), \ldots, X_l(\eta)$ таким образом, что к классу (подмножеству) $X_k(\eta)$ относятся все точки x, находящиеся к центру θ^k ближе, чем к любому другому. Интеграл $\int_{X_k(\eta)} \|x - \theta^k\|^2$ определяет рассеяние точек x в подмножестве $X_k(\eta)$.

Предположим, что на множестве X задано распределение $\mathrm{P}(\cdot)$. Определим функционал качества кластеризации по правилу η:

$$F(\eta) = \sum_{k=1}^{l(\eta)} \int_{X_k(\eta)} q^k(x, \eta) \mathrm{P}(dx).$$

Тогда задача кластеризации множества данных из l классов состоит в определении набора центров $\eta_\star^l$, минимизирующего суммарную стоимость разбиения. Задача устойчивой кластеризации является обобщением для случая нахождения заранее неизвестного оптимального значения количества классов $l_\star$ и соответствующего набора центров $\eta_\star^{l_\star}$.

Если в каком-то смысле можно было бы говорить о дифференцируемости функционала F, то искомый набор центров $\eta_\star$ должен был бы удовлетворять уравнению $\nabla F(\eta_\star) = 0$, которое можно было бы попытаться решить традиционными средствами. Сложность рассматриваемой задачи заключается в недифференцируемости функционала F. При известном распределении $\mathrm{P}(\cdot)$ сформулированные задачи все-таки могут быть точно или приближенно решены. Нетрудно убедиться, что в описанном выше примере множества $X_k(\eta)$ имеют вид многогранников, а минимизирующий функционал среднего риска набор центров $\eta_\star^{l_\star} = \left(\theta_\star^1, \theta_\star^2, \ldots, \theta_\star^{l^\star}\right)$ совпадает с *центрами тяжести* этих множеств, т. е.

$$\theta_\star^k = \frac{\int_{X_k(\eta_\star)} x \mathrm{P}(dx)}{\int_{X_k(\eta_\star)} \mathrm{P}(dx)},\ k = 1, 2, \ldots, l^\star.$$

Приведенные соображения отвечают интуитивному представлению о разбиении множества X на $l_\star$ непересекающихся классов, причём центры тяжести соседних множеств находятся на прямой, ортогональной разделяющей множества грани.

В системах, работающих в режиме реального времени в условиях изменяющейся со временем обстановки, распределение $\mathrm{P}(\cdot)$ очень часто бывает неизвестно. Будем предполагать, что в режиме реального времени на вход поступает последовательность $x_1, x_2, \ldots$ сигналов, порожденная неизвестным распределением $\mathrm{P}(\cdot)$. *Требуется* предложить алгоритм построения последовательности оценок $\{\hat{\eta}_n\}$ набора $\eta_\star$, минимизирующего определенный выше функционал среднего риска. Решение задачи дополнительно осложняется тем, что на практике функции $q^k(\cdot, \cdot),\ k = 1, 2, \ldots, l(\eta)$ не всегда заданы аналитически, но доступны измерению их значения (может быть

с помехами):

$$y^k(x,\eta) = q^k(x,\eta) + v^k,\ k = 1,2,\ldots,l(\eta).$$

В следующем пункте 1.3 описаны два примера задач распознавания образов: распознавание слов речи и распознавание печатных текстов на арабском языке.

В п. 1.4 рассматриваются основные программные средства аналитического ПО, поддерживающего методы классификации и кластеризации. Обосновывается необходимость создания нового средства для разработки и анализа алгоритмов.

Во **второй главе** "Рандомизированные алгоритмы кластеризации" предлагаются новые рандомизированные алгоритмы кластеризации и приведены результаты по исследованию их свойств.

В п. 2.1 для случая известного количества классов описывается метод k-средних, предложенный Дж.Хартиганом и М.Вонгом, широко применяемый при стандартных предположениях о свойствах погрешностей в измерениях, и рандомизированный алгоритм стохастической аппроксимации (РАСА), предложенный О.Н.Граничиным и О.А.Измаковой, работоспособный при ограниченных неконтролируемых (*unknown but bounded*) погрешностях.

В п. 2.2 формулируется Теорема 1 о свойствах оценок РАСА алгоритма в применении к задаче о дикторонезависимом распознавании образов слов речи. В предлагаемом новом способе звуковые сигналы последовательно преобразуются методом кепстральных коэффициентов (МКК) тоновой частоты в конечномерные вектора характеристических свойств размерности $n \approx 4000$, подаваемые впоследствии на вход алгоритма РАСА с двумя измерениями целевой функции на каждой итерации. Исходные элементы множества речевых сигналов принадлежат пространству с очень большой размерностью $n \approx 32000$, зависящей от частоты дискретизации аналого-цифрового преобразователя (АЦП).

МКК позволяет значительно сократить размерность фазового пространства, задавая отображение $\Phi : X \to \mathbb{R}^n$. Будем считать, что количество классов (различных слов) априори известно и равно l. Обозначим $\eta = \{l, (\theta^1, \theta^2, ..., \theta^l)\}$ и выберем для разных классов однотипные штрафные функции $q^k(x,\eta) = \|\Phi(x) - \theta^k\|^2$.

Зафиксируем некоторый начальный набор $\hat{\eta}_0 \in \mathbb{R}^{n\times l}$ и выберем последовательности положительных чисел $\{\alpha_i\}$ и $\{\beta_i\}$, стремящиеся к нулю. По алгоритму РАСА последовательность оценок $\{\hat{\eta}_i\}$ оптимального набора центров l классов $\eta_\star$ из пространства $\mathbb{R}^n$ строится следующим образом при помощи наблюдаемой последо-

вательности случайных независимых друг от друга векторов $\triangle_i \in \mathbb{R}^n$, $i = 1, 2, \ldots$, (называемых пробным одновременным возмущением и составленных из взаимно независимых бернуллиевских, равных ± 1, компонент):

$$\begin{cases} \widetilde{\eta}_i^{\pm} = \hat{\eta}_{i-1} \pm \beta_i \triangle_i J^T(x_i, \hat{\eta}_{i-1}), \\ \hat{\eta}_i = \mathcal{P}_\Theta \left(\hat{\eta}_{i-1} - \alpha_i J^T(x_i, \hat{\eta}_{i-1}) \frac{Y(x_i, \widetilde{\eta}_i^+) - Y(x_i, \widetilde{\eta}_i^-)}{2\beta_i} \triangle_i J^T(x_i, \hat{\eta}_{i-1}) \right), \end{cases}$$

в котором $J^T(x_i, \hat{\eta}_i)$ — l-мерный вектор, составленный из нулей и одной единицы, соответствующей координате с номером k в том случае, когда $\Phi(x_i)$ располагается ближе всего к множеству $X_k(\hat{\eta}_i)$; $Y(x_i, \widetilde{\eta}_i^{\pm}) = Q(x_i, \widetilde{\eta}_i^{\pm}) + V_i^{\pm}$ — l–мерные векторы, составленные из измеренных с помехами в соответствующих точках значений штрафных функций; $V_i^{\pm}$ — соответствующие вектора из ошибок наблюдений; $\mathcal{P}_\Theta$ – оператор проектирования на некоторое выпуклое замкнутое ограниченное подмножество $\Theta \subset \mathbb{R}^{n \times l}$, которое содержит $\eta_\star$. Условия состоятельности последовательности оценок $\{\hat{\eta}_n\}$ сформулированы в Теореме 1.

Теорема 1 *Пусть выполнены условия:* $|v_i| \le C_v$, $C_v > 0$;
$\forall i \ge 1$ *случайные вектора* $V_1^{\pm}, V_2^{\pm} \ldots, V_i^{\pm}$ *и* $x_1, x_2, \ldots, x_{i-1}$ *не зависят от* x_i, Δ_i, *случайный вектор* x_i *не зависит от* Δ_i;
$\sum_i \alpha_i = \infty$ *и* $\alpha_i \to 0$, $\beta_i \to 0$, $\alpha_i \beta_i^{-2} \to 0$ *при* $i \to \infty$;
расстояния в $\mathbb{R}^m$ *между образами разных классов более* $r_{\max} + 2C_v$, *т. е.*

$$\operatorname{dist}(\Phi(X_k(\eta_\star)), \Phi(X_j(\eta_\star))) \ge r_{\max} + 2C_v \quad \forall j \ne k,$$

где

$$r_{\max} = \max_{k \in 1:l} \max_{x \in X_k(\eta_\star)} \|\Phi(x) - \theta_\star^k\|^2.$$

Если для последовательности оценок $\{\hat{\eta}_i\}$, *доставляемых PACA при произвольном выборе* $\hat{\eta}_0$, *выполнено*

$$\overline{\lim_{i \to \infty}} J(x_i, \hat{\eta}_{i-1})^T Q(x_i, \hat{\eta}_{i-1}) \le r_{\max} + C_v,$$

тогда *последовательность оценок* $\{\hat{\eta}_i\}$ *сходится в среднеквадратичном смысле:* $\lim_{i \to \infty} \mathrm{E}\{\|\hat{\eta}_i - \eta_\star\|^2\} = 0$, *при* $i \to \infty$, *к одному из наборов* $\eta_\star$, *состоящему из векторов* $\theta_\star^1, \theta_\star^2, \ldots, \theta_\star^l$.

Если, более того, $\sum_i \alpha_i \beta_i^2 + \alpha_i^2 \beta_i^{-2} < \infty$, **то** $\hat{\eta}_i \to \eta_\star$ *при* $i \to \infty$ *с вероятностью*

единица.

Далее в п. 2.3 описываются несколько способов определения априори неизвестного количества кластеров в множестве данных X, основные идеи которых сводятся к заданию максимально возможной верхней границы $l_{\max}$ и вычислению некоторых индексных функций $Ind(l)$, характеризующих оптимальные разбиения множества X на l классов. Многие из таких функций строятся через степени рассеяния внутри класса, определяющие так называемые "искажения". В работе К.Сьюгер и Г.Джеймса обоснована целесообразность использования индексных функций вида $Ind(l) = I(D_l)$, где

$$D_l = \min_{i \in 1:l} \int_{X_i(l,\eta_\star^l)} q^i(x,(l,\eta_\star^l))\mathrm{P}(dx),$$

для анализа кривой "искажений" $D_l : [1, l_{max}] \to \mathbb{R}$, которая монотонно убывает с ростом l. В частности предлагается рассчитывать $I(D) = (D_l)^{-n/2}$. При этом трудоемкость алгоритмов прямопропорциональна размерности фазового пространства, выбору максимального количества кластеров $l_{\max}$ и среднему времени на кластеризацию при заданном l.

Существенное сокращение трудоемкости достигается при использовании описанного в п. 2.4 нового рандомизированного алгоритма достоверного определения количества кластеров с задаваемой вероятностью, который является развитием идей К.Сьюгер и Г.Джеймса.

Будем считать априори известными два положительных числа B и C:

$$B \geq |I(l_\star) - I(l_\star - 1)|,\ C \geq \max\nolimits_{j=2,\ldots,l_\star-1,l_\star+1,\ldots,l_{\max}} |I(l_j) - I(l_j - 1)|,$$

где $l_\star$ – оптимальное количество кластеров.

Зафиксируем три натуральных числа K, L, N и выберем из набора $1 : l_{\max}$ независимо с равномерным распределением K групп по L точек $\bar{i}_j$, $j = 1, \ldots, K$. Для каждого набора $\bar{i} \in 1 : l_{\max}$ обозначим $sp_{\bar{i}}(\cdot)$ равномерную аппроксимацию функции $I(\cdot)$ многочленами Чебышева степени небольшей N, построенную по соответствующим $\bar{i}$ точкам при условии выбора коэффициентов аппроксимации, непревосходящими C/K^2. Обозначим $\gamma = \max_{j \in 1:K} \max_{z \in \bar{i}_j} \left| sp_{\bar{i}_j}(z) - I(z) \right|$ – максимальную невязку. Определим на всем интервале $[1, l_{\max}]$ характеристическую функцию

$$\chi(z) = \max_{j \in 1:K} sp_{\bar{i}_j}(z) - \min_{j \in 1:K} sp_{\bar{i}_j}(z).$$

Теорема 2 *Если при выборе достаточно малых параметров уровня и достоверности* $\epsilon, \beta \leq 1$ *выполняется условие* $L \geq \frac{N+1}{\epsilon\beta} - 1$ *и множество*

$$Z = \{z : \chi(z) > 2(\gamma + 2(C + B)\epsilon)\}$$

не пустое, **тогда** *с вероятностью* $p = (1-\beta)^2$ *множество* Z *содержит точку* $l_\star$.

Доказательства теорем приведены в п. 2.5.

В **третьей главе** описана разработанная автором система для анализа и тестирования алгоритмов классификации и кластеризации. В п. 3.1 описана структура программного комплекса, п. 3.2 посвящен вопросам визуализации данных и методам снижения размерности, в п. 3.3 описана апробация алгоритмов устойчивой кластеризации.

При широком использовании кластеризации возникает необходимость в программных средствах, предоставляющих гибкие возможности для анализа данных и алгоритмов, а так же для удобного представления результатов. В рамках диссертационной работы была создана система, поддерживающая добавление, просмотр, редактирование и анализ данных и алгоритмов пользователя. Существует возможность генерировать код и сохранять результаты. Принцип работы системы схематично изображен на рис. 1.

Пользователь может выбрать сразу несколько алгоритмов и входных данных, которые поступают на обработку в вычислительное ядро. Результаты вычислений передаются анализатору, который обрабатывает результаты и возвращает их в структурированном виде. Алгоритмы можно задавать с помощью языка метаописания. Входные данные загружаются из базы данных либо генерируются искусственно. Кроме того есть возможность для некоторых алгоритмов использовать оракул, который заранее знает верный ответ и проводит верификацию результатов.

Система была реализована на языке Java с использованием технологий JSP, JDBC, JavaServlets, Xml-RPC, JavaScript, сервлет-контейнера Tomcat и сервера базы данных MySQL, а также с использованием вычислительного ядра, реализованного на C Sharp. Наличие такой системы позволяет структурировать уже известную информацию об алгоритмах классификации и кластеризации, а также исследовать новые эффективные методы.

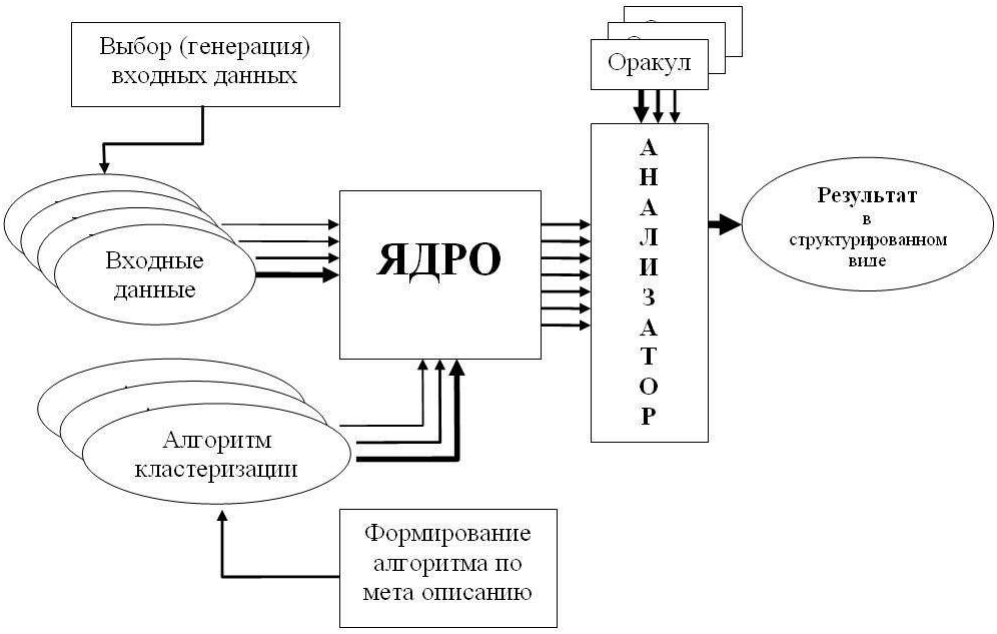

Рис. 1: Схема работы программного комплекса для разработки и анализа алгоритмов классификации и кластеризации.

В заключении диссертации подведены итоги проведенного и завершенного в рамках поставленных задач исследования.

Работы автора по теме диссертации

Статьи в журналах, рекомендованных ВАК:

[1] *Граничин О. Н., Шалымов Д. С.* Решение задачи автоматического распознавания отдельных слов речи при помощи рандомизированного алгоритма стохастической аппроксимации // Нейрокомпьютеры: разработка, применение. – М., Радиотехника, 2009. – N 3. – С. 58-64.

[2] *Шалымов Д. С.* Распознавание слитной речи с использованием рандомизированного алгоритма стохастической аппроксимации // Вестник СПбГУ. Сер. 10: Прикладная математика, информатика, процессы управления. - С.: Изд-во СПбГУ, 2009. – N 3. – С. 171-181.

[3] *Шалымов Д. С.* Рандомизированный метод определения количества кластеров на множестве данных // Научно-технический вестник СПбГУ ИТМО. - Санкт-Петербург, 2009. – N 5. – С. 111-116.

Другие публикации:

[4] *Шалымов Д. С.* Рандомизированный алгоритм стохастической аппроксимации в задаче распознавания отдельных слов речи // В сб. "Стохастическая оптимизация в информатике" под ред. О. Н. Граничина, Вып. 2. – Изд-во С.-Петерб. ун-та, 2006. – С. 207-218.

[5] *Granichin O. N., Shalymov D. S.* New breed stochastic hybrid computers // In: Proc. of the 3-rd Int. IEEE Scientific Conf. on Physics and Control (PhysCon 2007). – Potsdam, Germany, 2007. – С. 178.

[6] *Граничин О. Н., Шалымов Д. С.* Новые компьютеры. Вычислительные устройства будущего // Компьютерные инструменты в образовании. – М., 2007. – N 6. – С. 23-31.

[7] *Шалымов Д. С.* Автоматическое распознавание печатных текстов арабского языка // В сб. "Стохастическая оптимизация в информатике" под ред. О. Н. Граничина. Вып. 3. – Изд-во С.-Петерб. ун-та, 2007. – С. 124-137.

[8] *Шалымов Д. С.* Методы стохастической оптимизации в задаче распознавания печатных текстов арабского языка // В сб. трудов пятой межд. научно-практической конф. "Исследование, разработка и применение высоких технологий в промышленности". – Санкт-Петербург, 2008. – С. 140-142.

[9] *Шалымов Д. С.* Дикторонезависимое распознавание речи при помощи рандомизированного алгоритма стохастической аппроксимации // В сб. трудов пятой межд. научно-практической конф. "Исследование, разработка и применение высоких технологий в промышленности". – Санкт-Петербург, 2008. – С. 131-138.

[10] *Shalymov D. S.* Noise Robust Isolated Words Recognition Problem Solving Based On Simultaneous Perturbation Stochastic Approximation Algorithm // The 20th Int. Conf. "Continuous Optimization and Knowledge-Based Technologies". – Neringa, Lithuania, 2008. – С. 112-118.

[11] *Granichin O. N., Shalymov D. S.* Speaker-Independent Isolated Words Recognition Problem Solving Based On Simultaneous Perturbation Stochastic Approximation Algorithm // Yalta Conf. on Discrete and Global Optimization. – Yalta, Ukraine, 2008. – С. 13.

[12] *Граничин О. Н., Шалымов Д. С.* Исследование и рандомизация алгоритмов устойчивой кластеризации на основе индексов // Сб. трудов III Межд. научно-

практической конф. “Современные информационные технологии и ИТ-образование”. – М., 2008. – Режим доступа: http://2008.it-edu.ru/pages/Conference-works, свободный.

[13] *Шалымов Д. С.* Алгоритмы устойчивой кластеризации на основе индексных функций и функций устойчивости // В сб. “Стохастическая оптимизация в информатике” под ред. О. Н. Граничина. Вып. 4. – Изд-во С.-Петерб. ун-та, 2008. – С. 236-248.

[14] *Шалымов Д. С.* Рандомизированный алгоритм стохастической аппроксимации в задаче распознавания печатных текстов арабского языка // VIII Межд. конф. “Идентификация систем и задачи управления” (System Identification and Control Problems (SICPRO ’09)). – М., CD Proceedings, ISBN 978-5-91450-024-2, 2009.

[15] *Шалымов Д. С.* Определение количества кластеров на множестве данных с использованием рандомизированных алгоритмов // В сб. трудов XI конф. молодых ученых “Навигация и управление движением”. – Санкт-Петербург, 2009. – Режим доступа: http://www.elektropribor.spb.ru/cnf/kmu11/rrefs.html, свободный.

[16] *Shalymov D., Skrygan K., Lyubimov D.* Clustering Algorithms Meta Applier (CAMA) Toolbox // SYRCoSE (Spring Young Researchers Colloquium on Software Engineering). – Moscow, 2009. – С. 61-64.

[17] *Шалымов Д. С.* Рандомизированные алгоритмы в задаче кластеризации данных // В сб. трудов Первой традиционной всероссийской молодежной летней школе “Управление, информация и оптимизация”. – Переславль-Залесский, 2009. – С. 25-31.

[18] *Шалымов Д. С.* On-line кластеризация данных с использованием рандомизированных алгоритмов // Сб. трудов VI школы-семинара молодых ученых “Управление большими системами”. - Ижевск, 2009. – С. 389-399.

Подписано в печать 14.10.2009 г.
Формат бумаги 60 x 84 1/16. Бумага офсетная.
Печать цифровая. Усл. печ. л. 1,0. Тираж 100 экз. Заказ 4521.
Отпечатано в отделе оперативной полиграфии химического факультета СПбГУ.
198504, Санкт-Петербург, Петродворец, Университетский пр. 26.