

На правах рукописи

СЕРГИЕНКО РОМАН БОРИСОВИЧ

**АВТОМАТИЗИРОВАННОЕ ФОРМИРОВАНИЕ НЕЧЕТКИХ
КЛАССИФИКАТОРОВ САМОНАСТРАИВАЮЩИМИСЯ
КОЭВОЛЮЦИОННЫМИ АЛГОРИТМАМИ**

05.13.17 - Теоретические основы информатики

АВТОРЕФЕРАТ
диссертации на соискание ученой степени
кандидата технических наук

Красноярск – 2010

Работа выполнена в ФГОУ ВПО «Сибирский государственный аэрокосмический университет имени академика М.Ф. Решетнёва»,
г. Красноярск

Научный руководитель: доктор технических наук, профессор
Семенкин Евгений Станиславович

Официальные оппоненты: доктор технических наук, профессор
Миркес Евгений Моисеевич

кандидат технических наук
Цой Юрий Робертович

Ведущая организация: Институт проблем управления РАН
им. В.А. Трапезникова (г. Москва)

Защита состоится «10» декабря 2010 г. в 14 часов на заседании диссертационного совета Д 212.099.11 при Сибирском федеральном университете по адресу 660074, г. Красноярск, ул. Киренского, 26, ауд. УЛК 115.

С диссертацией можно ознакомиться в библиотеке Сибирского федерального университета по адресу: г. Красноярск, ул. Киренского, 26, ауд. Г 2-74.

Автореферат разослан «29» октября 2010 г.

Ученый секретарь
диссертационного совета,
кандидат технических наук,
доцент



Покидышева Л.И.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность. На сегодняшний день интенсивно развивается направление, связанное с интеллектуализацией методов обработки и анализа данных. Интеллектуальные системы анализа данных (ИСАД) призваны минимизировать усилия лица, принимающего решения (ЛПР), в процессе анализа данных, а также в настройке алгоритмов анализа. Многие ИСАД позволяют не только решать классические задачи принятия решения, но и способны выявлять причинно-следственные связи, скрытые закономерности в системе, подвергаемой анализу.

Одна из наиболее распространенных задач анализа данных – классификация. К задачам классификации сводятся диагностирование заболеваний в медицине, принятие решения о выдаче кредита клиенту банка, распознавание изображений и звука и многие другие задачи. На сегодняшний день создано большое количество разнообразных алгоритмов классификации (в частности, различные алгоритмы автоматической классификации разрабатываются в научной школе Дорофеюка А.А. в Институте проблем управления РАН), разработаны формализованные теории конструирования высокоэффективных композиций из этих алгоритмов (например, метод алгоритмических композиций научной школы академика РАН Журавлёва Ю.И.). Однако недостатком большинства из существующих алгоритмов классификации является работа по принципу «чёрного ящика» – невозможность явной интерпретации закономерностей, приводящих к отнесению объекта классификации к тому или иному классу.

Такого недостатка лишены методы вербального анализа решений, разработкой которых, в частности, занимаются в Институте системного анализа РАН, а также нечеткий классификатор, являющийся разработкой японских специалистов в области нечетких систем во главе с Х. Ишибучи. Нечеткий классификатор представляет собой базу нечетких правил. Каждое нечеткое правило – выражение причинно-следственной закономерности отнесения объекта к какому-либо классу в лингвистической форме, доступной для прямого восприятия экспертом. Таким образом, нечеткий классификатор представляет собой одновременно алгоритм классификации и инструмент интеллектуального анализа данных (Data Mining) для обнаружения скрытых знаний и закономерностей.

Формирование нечеткого классификатора сводится к задаче оптимизации – выбору наилучшей базы нечетких правил из множества существующих. Данная задача оптимизации отличается существенной вычислительной и алгоритмической сложностью, обусловленной высокой размерностью решаемой задачи оптимизации, процедурно заданной целевой функцией (вычисление которой, как правило, трудоёмко), наличием дискретных переменных и др. В такой ситуации использование классических методов оптимизации неприемлемо. Для задачи формирования нечеткого классификатора целесообразно использовать эволюционные алгоритмы, которые хорошо зарекомендовали себя при решении именно такого рода задач оптимизации.

Однако применение эволюционных алгоритмов оптимизации сопряжено с другой серьезной проблемой – высокой сложностью и трудоемкостью настройки алгоритмов на решаемую задачу в связи с большим числом возможных комбинаций их настроек (селекции, мутации, скрещивания и некоторых других). Эффективность одной и той же настройки на разных задачах и различных настройках на одной и той же задаче может изменяться в очень широком диапазоне. Поэтому выбор настроек наугад является неприемлемым, так как многие комбинации параметров алгоритма оказываются неработоспособными, а тщательная настройка под новую задачу является чрезмерно трудоемкой, особенно для таких вычислительно сложных задач, как формирование нечеткого классификатора. Для разрешения указанной проблемы предлагается использование так называемого коэволюционного подхода для адаптации стратегии оптимизации, позволяющего отказаться от выбора настроек эволюционного алгоритма за счет взаимодействия (конкуренции и кооперации) множества эволюционных алгоритмов с различными настройками.

Таким образом, разработка и исследование самонастраивающихся коэволюционных алгоритмов и дальнейшее их использование в системах формирования нечетких классификаторов для одновременного решения задач классификации и интеллектуального анализа данных с выявлением скрытых знаний, является **актуальной** научно-технической задачей.

Целью диссертационной работы является повышение эффективности, быстродействия и адаптивности эволюционного метода формирования нечеткого классификатора с использованием самонастраивающихся коэволюционных алгоритмов.

Сформулированная цель предопределила следующую совокупность решаемых задач:

1. Выполнить аналитический обзор существующих методов классификации и оптимизации с целью установления целесообразности использования нечеткого классификатора, эволюционных алгоритмов как метода формирования нечетких систем, коэволюционных алгоритмов как инструмента автоматизации выбора настроек эволюционного алгоритма.

2. Провести комплексные исследования эффективности коэволюционного подхода как инструмента автоматизации выбора настроек эволюционного алгоритма на различных классах задач.

3. Разработать метод формирования нечетких классификаторов с использованием коэволюционных алгоритмов, обладающий преимуществами перед существующими методами.

4. Реализовать разработанные алгоритмы в виде программных систем и апробировать их на реальных практических задачах.

Методы исследования. При выполнении диссертационной работы использовались методы теории оптимизации, теории вероятностей и математической статистики, анализа данных, методы снижения размерности, теории эволюционных алгоритмов, теории нечетких систем; методики разработки и исследования моделей и алгоритмов анализа данных, выявления

закономерностей в данных и их извлечения, применения бионических принципов и моделей в информационных технологиях.

Научная новизна результатов диссертационной работы состоит в следующем:

1. Разработан новый самонастраивающийся коэволюционный алгоритм условной оптимизации, отличающийся от известных самонастройкой метода учёта ограничений и пропорционально-групповым способом реализации модифицированного оператора миграции.

2. Разработан новый самонастраивающийся коэволюционный алгоритм многокритериальной оптимизации, отличающийся от известных отсутствием независимой работы индивидуальных алгоритмов в коэволюционной комбинации и способом оценивания эффективности индивидуальных алгоритмов.

3. Разработан новый метод формирования нечеткого классификатора, отличающийся от известных схемой комбинирования Мичиганского и Питтсбургского подходов, наличием формализованной процедуры отбора стартовых нечетких правил и использованием самонастраивающихся коэволюционных алгоритмов.

Теоретическая значимость результатов диссертационного исследования состоит в том, что разработаны, исследованы и апробированы новые коэволюционные алгоритмы для автоматизации выбора настроек генетических алгоритмов и новый метод формирования нечеткого классификатора самонастраивающимися коэволюционными алгоритмами для решения задач классификации и интеллектуального анализа данных, что является существенным вкладом в теорию и практику исследования методов преобразования информации в данные и знания, методов машинного обучения и обнаружения новых знаний.

Практическая ценность. Разработанные в ходе исследования алгоритмы и методы реализованы в виде двух программных систем: «Коэволюционный алгоритм условной многокритериальной оптимизации» и «Нечеткий коэволюционный классификатор». Первая программная система предназначена для решения различных задач оптимизации. Программная система использовалась для решения следующих практических задач: формирование оптимального инвестиционного портфеля предприятия, формирование оптимального кредитного портфеля банка, планирование расписания посадок самолётов.

Вторая программная система предназначена для решения задач классификации с одновременным выявлением лингвистических знаний в соответствующих проблемных областях. Программная система апробировалась при решении практических задач диагностики заболевания печени, выдачи банковского кредита физическим лицам, классификации видов ириса, распознавания спутниковых изображений и некоторых других.

Реализация результатов работы. Программные системы использовались в качестве лабораторных установок студентами Института информатики и

телекоммуникаций СибГАУ по направлению «Системный анализ и управление» при выполнении курсовых работ. Результаты решения программной системой «Нечеткий коэволюционный классификатор» задачи распознавания спутниковых изображений использовались в рамках проекта «Студенческий спутник СибГАУ» в сотрудничестве с Центром управления полетами СибГАУ.

Полученные результаты применялись в ходе выполнения работ по созданию систем управления сложными механическими системами (двойной перевернутый маятник) во время стажировки в Высшей технической школе г. Ульм (Hochschule Ulm), ФРГ, 2008 г.

Диссертационная работа поддержана Фондом содействия развитию малых форм предприятий в научно-технической сфере по программе «У.М.Н.И.К.» («Участник молодежного научно-инновационного конкурса») в рамках НИОКР «Разработка коэволюционного вероятностного алгоритма для автоматизированного проектирования интеллектуальных информационных технологий» на 2008-2011 гг. Работа финансировалась из средств госбюджета в рамках НИР Б1.01.05 «Разработка и исследование бионических методов идентификации и оптимизации сложных систем» ЕЗН СибГАУ, а также в рамках выполнения проекта «Система поддержки принятия решения при проектировании интегрированных систем безопасности», ставшего победителем конкурса инновационных проектов СибГАУ на 2007-2008 гг. Работа поддержана грантом в рамках конкурса научно-технического творчества молодежи города Красноярска «НТТМ-2010». Диссертационное исследование удостоено диплома 1 степени Российской ассоциации искусственного интеллекта.

Диссертационное исследование проводилось также в рамках НИР № 2.1.1./2710 «Математическое моделирование инвестиционного развития региональных экономических систем» АВЦП «Развитие научного потенциала высшей школы (2009–2010 годы)» и НИР НК-136П/3 "Автоматизированная система решения сложных задач глобальной оптимизации многоагентными стохастическими алгоритмами" ФЦП «Научные и научно-педагогические кадры инновационной России» на 2009 – 2013 годы.

Основные защищаемые положения:

1. Самонастраивающийся коэволюционный алгоритм безусловной однокритериальной оптимизации обеспечивает эффективность, не уступающую стандартному генетическому алгоритму с настройками, обеспечивающими среднюю эффективность.

2. Эффективность самонастраивающегося коэволюционного алгоритма условной однокритериальной оптимизации с использованием разработанной пропорционально-групповой схемы миграции превосходит эффективность стандартного генетического алгоритма с наилучшими настройками.

3. Разработанный новый метод формирования нечетких классификаторов самонастраивающимися коэволюционными алгоритмами позволяет получать эффективную базу знаний с небольшим числом правил.

4. Нечеткий классификатор, формируемый разработанным методом, превосходит по точности классификации другие известные подходы.

Публикации. По теме диссертации опубликовано двадцать три работы, в том числе четыре в изданиях из перечня ВАК, а также зарегистрированы три программные системы.

Апробация работы.

Результаты диссертационной работы были доложены и обсуждены на Всемирном конгрессе по вычислительному интеллекту (IEEE World Congress on Computational Intelligence (WCCI'2010), Barcelona, Spain, 2010); Национальной конференции по искусственному интеллекту с международным участием КИИ-2010 (Тверь, 2010); Всероссийской научной конференции «Теория и практика системного анализа» (Рыбинск, Институт системного анализа РАН и РГАТА 2010); Международных научно-технических конференциях «Интеллектуальные системы» AIS'08 и AIS'09 (Дивноморское, 2008, 2009), на Всероссийских научно-практических конференциях с международным участием «Информационные технологии и математическое моделирование» (Томский университет, 2007, 2009), Конференции-конкурсе «Технологии Microsoft в теории и практике программирования» (Томск, ТПУ, 2010), Всероссийских конференциях молодых ученых по математическому моделированию и информационным технологиям (Красноярск, ИВМ СО РАН, 2006, и Новосибирск, ИВТ СО РАН, 2007), на Международных научно-практических конференциях «Решетневские чтения» (Красноярск, СибГАУ, 2006-2009), на научном семинаре в Институте проблем управления РАН (Москва, 2010), а также на ряде молодежных и студенческих конференций.

Структура работы. Диссертация содержит 152 страницы основного текста и состоит из введения, четырёх глав, заключения, списка литературы на 157 источников и 8 приложений, содержащих 23 страницы, основной текст диссертации включает 47 таблиц, 26 рисунков.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

Во **введении** обоснована актуальность работы, сформулирована цель и поставлены задачи исследования, рассмотрены вопросы научной новизны и практической ценности проведенных исследований, изложены основные положения, выносимые на защиту.

Первая глава посвящена обзору существующих алгоритмов классификации и оптимизации с целью установления целесообразности использования нечеткого классификатора как алгоритма классификации и инструмента интеллектуального анализа данных, эволюционных алгоритмов как метода формирования нечетких систем, коэволюционных алгоритмов как инструмента автоматизации выбора настроек эволюционного алгоритма.

Рассмотрены различные современные методы глобальной оптимизации, как детерминированные, так и стохастические: мультистарт алгоритмов локального поиска, алгоритм группировок, методы покрытий, интервалов, траекторные методы, метод туннелей, метод имитации обжига, усреднения координат, алгоритмы псевдобулевой оптимизации, эволюционные алгоритмы и др. Выяснено, что преимущество эволюционных алгоритмов перед другими

методами оптимизации заключается в способности эффективно решать многомерные, многоэкстремальные задачи; при зашумлённости целевой функции, её неявном (например, алгоритмическом) задании; при дискретности переменных. Важнейший подкласс эволюционных алгоритмов – *генетические алгоритмы* (ГА).

Далее раскрывается такое понятие, как коэволюционный генетический алгоритм. Под *коэволюционными алгоритмами* понимается система взаимодействующих между собой генетических алгоритмов (иными словами, множества популяций индивидов-решений генетического алгоритма). Таким образом, коэволюционный алгоритм является представителем мультиалгоритмического подхода (Ю.И. Журавлев, Л.А. Растринин) в классе эволюционных алгоритмов.

Все схемы коэволюционных алгоритмов объединяет два общих свойства:

1) в коэволюционном алгоритме обязательно представлено множество эволюционирующих в той или иной степени изолированно друг от друга подпопуляций;

2) подпопуляции коэволюционного алгоритма тем или иным способом взаимодействуют друг с другом.

Кроме того, алгоритмы, составляющие коэволюционную комбинацию обладают, как правило, различными свойствами или характеристиками. В противном случае речь идет лишь о мультипопуляционных эволюционных алгоритмах.

В настоящее время к собственно коэволюционным алгоритмам принято относить две схемы, получившие названия *«кооперативной»* (*cooperative coevolution*) и *«конкурирующей»* (*competitive coevolution*) коэволюции соответственно. В кооперативном коэволюционном алгоритме осуществляется декомпозиция решаемой задачи оптимизации (множества параметров и/или целевой функции), что упрощает решение вычислительно и структурно сложных задач. Под конкурирующими коэволюционными алгоритмами понимается сразу несколько различных подходов: взаимодействие двух подпопуляций по схеме «хозяин-паразит», метод конкурирующих подпопуляций, отличающихся поисковой областью и метод конкурирующих подпопуляций, отличающихся поисковой стратегией (настройками параметров алгоритма).

Последний из перечисленных методов служит для решения проблемы ручного выбора настроек генетического алгоритма, т.к. перераспределение ресурсов между алгоритмами в ходе конкуренции автоматически обеспечивает доминирование алгоритма (субпопуляции) с наилучшими настройками. В дальнейшем данный метод был модернизирован путем введения взаимодействия между субпопуляциями за счет миграции лучших индивидов.

Обобщенная схема «конкурентно-кооперативного» коэволюционного алгоритма для автоматизации выбора настроек ГА:

- выбор индивидуальных алгоритмов;

- задание параметров коэволюционного алгоритма (размер общего ресурса, величина интервала адаптации, размер штрафа «проигравшего» алгоритма, размер «социальной карты» - минимально допустимого размера подпопуляций);
- независимая работа выбранных алгоритмов в течение интервала адаптации (обычно около 5 поколений);
- оценка алгоритмов;
- перераспределение ресурсов;
- миграция лучших индивидов во все подпопуляции.

Далее рассматриваются методы условной оптимизации и методы многокритериальной оптимизации в генетических алгоритмах.

Методы учета ограничений в ГА для решения задач условной оптимизации: метод «смертельных» штрафов, метод статических штрафов, метод динамических штрафов, метод адаптивных штрафов, «лечение» локальным поиском, метод поведенческой памяти, коэволюционный алгоритм поиска множителей Лагранжа и некоторые другие.

Методы многокритериальной оптимизации в ГА: VEGA (Vector Evaluated Genetic Algorithm), FFGA (Fonseca and Fleming's Multiobjective Genetic Algorithm), NPGA (Niche Pareto Genetic Algorithm), группа методов NSGA (Non-Dominated Sorting Genetic Algorithm), метод SPEA (Strength Pareto Evolutionary Algorithm), а также его модификация SPEA 2. По результатам предыдущих исследований показано, что в большинстве случаев SPEA 2 превосходит по эффективности остальные методы.

Рассматриваются методы классификации: решающее Байесовское правило, различные статистические и детерминистские методы классификации, методы конструирования высокоэффективных композиций из алгоритмов распознавания. Выяснено, что большинство из них работают по принципу «чёрного ящика» с невозможностью явной интерпретации причинно-следственных связей отнесения объекта к тому или иному классу.

Такого недостатка лишены методы вербального анализа решений (в частности, методы ЦИКЛ, ИСКРА, ПАКС, разработанные в Институте системного анализа РАН А. А. Асановым, О. И. Ларичевым, Г. В. Ройзензоном), а также нечеткий классификатор, разработанный группой японских специалистов в области нечетких систем (Х. Ишибучи, Т. Накашима, Т. Курода, Т. Мурата). Нечеткий классификатор – алгоритм классификации, основанный на извлечении нечетких правил из массивов данных. Каждое правило в нечетком классификаторе содержит значения лингвистических термов для всех информативных признаков (включая терм «игнорирования»), номер класса, которому соответствует данное правило, а также уровень значимости в интервале $[0;1]$, показывающий степень достоверности правила. Нечеткий классификатор относится и к алгоритмам классификации, и к технологиям интеллектуального анализа данных (Data Mining) для обнаружения скрытых знаний и закономерностей.

Для автоматического формирования нечеткого классификатора целесообразно использовать эволюционные алгоритмы. Существует два

основных подхода к конструированию нечетких систем эволюционными алгоритмами – Мичиганский (индивид – одно правило) и Питтсбургский (индивид – база правил в целом), каждый из которых обладает своими недостатками и преимуществами. Является актуальной задачей разработка схемы формирования нечеткого классификатора с комбинированием Мичиганского и Питтсбургского подходов и с автоматической настройкой параметров ГА с использованием коэволюционного подхода.

Вторая глава посвящена исследованию эффективности коэволюционного подхода как инструмента автоматизации выбора настроек ГА на различных классах задач оптимизации. Исследование эффективности коэволюционного алгоритма (с полным перебором всех параметров ГА и набором статистики) собственно на задачах формирования нечеткого классификатора нецелесообразно ввиду очень значительных вычислительных затрат при решении таких задач. Поэтому выбраны тестовые задачи, а полученные результаты по аналогии переносятся на более сложные задачи.

Для задач безусловной однокритериальной оптимизации выбраны восемь тестовых задач размерности 2, традиционно считающиеся сложными для классических методов оптимизации: функции Растригина, Сомбреро, Шекеля, Розенброка, Гриванка, Катковника, аддитивная и мультипликативная потенциальная функции. Для задач условной однокритериальной оптимизации сконструированы десять тестовых задач (восемь с ограничениями-неравенствами, две с ограничениями-равенствами) с линейными и нелинейными (полиномиальными, тригонометрическими) целевой функцией и ограничениями. Также были взяты для тестирования две практические задачи на основе реальных данных: задача формирования оптимального инвестиционного портфеля предприятия и задача формирования оптимального кредитного портфеля банка. Первая задача состоит в составлении такого портфеля инвестиционных проектов, который приносит инвестору наибольшую прибыль. При этом должны выполняться ограничения по выделяемым средствам, норме прибыли и общей рискованности портфеля. Вторая задача состоит в формировании оптимального кредитного портфеля при наличии жестких ограничений по суммам имеющихся в наличии свободных кредитных ресурсов, их стоимости, процентным ставкам на выдаваемые кредиты, срокам привлечения ресурсов, максимальному размеру кредита на одного заемщика. Для многокритериальной оптимизации аналогично были сконструированы тестовые задачи, а также взяты упомянутые инвестиционная и банковская задачи в многокритериальной постановке.

При исследованиях используются три типа селекции (пропорциональная, турнирная с турниром 2, ранговая), три типа скрещивания (одноточечное, двухточечное, равномерное) и три типа метода учета ограничений для условной оптимизации (метод «смертельных» штрафов в комбинации с «лечением» локальным поиском, метод динамических штрафов, метод адаптивных штрафов). Используется адаптивная мутация, в многокритериальной оптимизации используется единственный метод SPEA2. Таким образом, число различных алгоритмов равно 9 для безусловной оптимизации и 27 для условной

оптимизации. Исследование эффективности коэволюционного алгоритма заключается в тестировании всех различных стандартных ГА и коэволюционного алгоритма, включающего все различные алгоритмы. При этом ресурсы (размер популяции и число поколений) для стандартного ГА и коэволюционного алгоритма совпадают. Главный критерий эффективности – надежность отыскания истинного оптимума с заданной точностью при многократном запуске, второй критерий – среднее число поколений отыскания наилучшего решения, характеризующий скорость сходимости алгоритма. При исследованиях проводился статистический анализ при многократных запусках с использованием рангового критерия Вилкоксона. В многокритериальных задачах критерии – доля истинно Паретовских точек и средний разброс в критериальном пространстве.

При разработке коэволюционного алгоритма условной оптимизации оказалось невозможным использование стандартной схемы миграции. Дело в том, что миграция основана на группировке индивидов из всех подпопуляций, сортировке их по пригодности и «раздаче» всем алгоритмам наилучших индивидов. Использование же различных методов учёта ограничений перестаёт делать пригодность единым критерием оценивания индивидов из различных подпопуляций. Были разработаны три модифицированных схемы миграции: пропорционально-групповая, паретовская и миграция с разделением по допустимости. По результатам исследований безоговорочно лучшей на всех задачах оказалась *пропорционально-групповая схема миграции*.

Адаптация коэволюционного алгоритма на класс задач многокритериальной оптимизации также была сопряжена с существенными изменениями. Алгоритм SPEA 2 направлен на поиск наиболее репрезентативной аппроксимации множества Парето (т.е. является методом сужения множество Парето). Наличие изолированных друг от друга в течение интервала адаптации подпопуляций в стандартном коэволюционном алгоритме приводит, во-первых, к высокой вероятности появления повторяющихся или близких друг к другу индивидов и, во-вторых, к появлению значительного числа индивидов, доминируемых индивидами из других подпопуляций. Поэтому было принято решение о наличии единой популяции для всех индивидуальных алгоритмов. При этом «закрепленная» за каждым алгоритмом часть потомков генерируется согласно настройкам этого алгоритма. Формирование новой популяции (из родителей и потомков) в каждом поколении осуществляется на основе результатов работы всех алгоритмов.

Далее приведены результаты сравнения наихудшего, среднего, наилучшего ГА и коэволюционного алгоритма на различных классах задач оптимизации (над. – надежность, пок. – среднее число поколений для нахождения наилучшего решения, ДП – доля истинно Паретовских решений, Р – разброс в критериальном пространстве, используется при равенстве ДП) для тестовых задач.

На основании проведенных исследований эффективности коэволюционного алгоритма были сделаны следующие обобщающие выводы.

Козэволюционный алгоритм решения задач безусловной однокритериальной оптимизации обеспечивает надежность отыскания оптимума не хуже среднестатистического стандартного ГА (в некоторых случаях надежность коэволюционного алгоритма может превосходить надежность наилучшего стандартного ГА) и всегда обеспечивает существенное сокращение времени сходимости по сравнению со стандартным ГА при любых настройках.

Козэволюционный алгоритм решения задач условной однокритериальной оптимизации на большинстве задач обеспечивает надежность отыскания оптимума выше или не хуже, чем стандартный ГА с наилучшими настройками, и обеспечивает существенное сокращение времени сходимости по сравнению со стандартным ГА при любых настройках.

Козэволюционный алгоритм решения задач многокритериальной оптимизации обеспечивает эффективность на уровне среднего алгоритма и всегда уступает наилучшему алгоритму.

Таблица 1

Исследования на задачах безусловной однокритериальной оптимизации

Задача	Наихудший ГА		Средний ГА		Наилучший ГА		Козэволюция	
	Над.	Пок.	Над.	Пок.	Над.	Пок.	Над.	Пок.
1	0,79	74,6	0,96	60,9	1,00	50,7	1,00	25,8
2	0,22	76,7	0,25	78,9	0,32	83,6	0,24	40,2
3	0,26	71,5	0,57	69,0	0,78	65,3	0,66	32,9
4	0,16	68,0	0,55	69,0	0,81	70,9	0,87	49,6
5	0,84	71,2	0,88	65,5	0,93	67,1	0,89	30,9
6	0,41	74,1	0,87	63,7	1,00	52,7	0,97	28,0
7	1,00	57,2	1,00	45,0	1,00	38,3	1,00	27,0
8	1,00	48,1	1,00	43,1	1,00	39,7	1,00	26,3

Таблица 2

Исследования на задачах условной однокритериальной оптимизации

Задача	Наихудший ГА		Средний ГА		Наилучший ГА		Козэволюция	
	Над.	Пок.	Над.	Пок.	Над.	Пок.	Над.	Пок.
1	0,016	69,8	0,249	64,1	0,736	70,8	0,990	57,8
2	0,014	41,3	0,661	63,5	1,000	55,2	1,000	23,0
3	0,012	15,0	0,679	62,1	1,000	60,6	1,000	31,4
4	0,190	56,0	0,735	59,6	1,000	62,6	1,000	27,0
5	0,000	-	0,066	73,2	0,292	73,4	0,990	48,0
6	0,014	57,2	0,550	61,5	1,000	68,0	0,982	39,4
7	0,000	-	0,080	70,5	0,320	71,0	0,996	50,2
8	0,000	-	0,420	78,8	0,748	70,0	1,000	44,8
9	0,050	15,0	0,294	51,4	0,568	71,8	0,802	54,0
10	0,028	16,5	0,172	50,8	0,372	75,4	0,340	59,0

Таблица 3

Исследования на задачах многокритериальной оптимизации

Алгоритм	Зад. 1 (ДП)	Зад. 2 (ДП)	Зад. 3 (ДП)	Зад. 4 (ДП)	Зад. 5 (ДП)	Зад. 6 (Р)	Зад. 7 (ДП)
Наихудший ГА	0,217	0,569	0,788	0,566	0,960	3,28	0,818
Средний ГА	0,246	0,611	0,807	0,582	0,973	3,34	0,830
Наилучший ГА	0,284	0,644	0,823	0,613	0,980	3,43	0,854
Козэволюция	0,242	0,620	0,810	0,573	0,972	3,32	0,830

Таким образом, эффективность коэволюционного подхода для решения проблемы выбора настроек ГА в большей мере проявляется на задачах однокритериальной оптимизации (в особенности, на задачах условной однокритериальной оптимизации) и в меньшей мере на задачах многокритериальной оптимизации. Кроме того, проведены исследования, показывающие эффективность использования коэволюционного алгоритма однокритериальной оптимизации в многопроцессорных вычислительных системах.

Поэтому для использования преимуществ коэволюционного алгоритма в задаче формирования нечеткого классификатора следует формулировать данную задачу в однокритериальной постановке, или в виде комбинаций однокритериальных задач. Помимо возможности использования преимуществ коэволюционного подхода, однокритериальная постановка требует меньших вычислительных ресурсов, обладает меньшей алгоритмической сложностью, позволяет эффективно использовать возможности многопроцессорных систем. Для таких вычислительно затратных задач, как формирование нечеткого классификатора, повышение эффективности вычислительной процедуры оказывается важнее преимуществ многокритериального подхода.

В **третьей главе** рассматривается разработка и исследование схемы формирования нечеткого классификатора с использованием самонастраивающихся коэволюционных алгоритмов.

Предлагаемая схема формирования нечеткого классификатора включает три основных этапа (помимо этапа фаззификации информативных признаков, осуществляемой тривиальным способом – равномерное заполнение нечеткими числами интервалов варьирования признаков в случае численных признаков или фаззификация дискретных значений в случае ранговых, качественных или бинарных признаков): формализованная процедура отбора стартовых правил с использованием априорной информации из обучающей выборки, этап улучшения стартовых правил Мичиганским методом (задача однокритериальной безусловной оптимизации), этап сокращения найденного множества правил Питтсбургским методом (задача однокритериальной условной оптимизации). Рассмотрим этапы более подробно.

Формирование начальной популяции для Мичиганского этапа (отбор стартовых правил). Пусть n – число нечетких правил, необходимое для заполнения популяции для Мичиганского этапа, k – число классов, l – число информативных признаков в задаче классификации. Последовательность шагов:

1) вычислить $m = \text{int}(n/k)$;

2) отсортировать элементы обучающей выборки по номеру класса в порядке возрастания и определить границы для каждого класса в массиве;

3) от $i:=1$ до k

от $j:=1$ до m

выполнить случайный выбор элемента класса i из обучающей выборки;

от $t:=1$ до l

определить ближайший центр нечеткого числа по признаку t ;
назначить соответствующее нечеткое число в нечеткое правило;
изменить нечеткое число на терм «игнорирование» с вероятностью 0,33;

4) заполнить популяцию сгенерированными нечеткими правилами.

Мичиганский этап формирования нечеткого классификатора. Индивиды представляют собой отдельные нечеткие правила. Длина хромосомы равна числу информативных признаков, каждый ген – число от 1 до 6, соответствующее нечеткому числу (для каждого признака используется пять значащих нечетких чисел, а также терм «игнорирования» признака). Функция пригодности индивидов – доверительный уровень правила, вычисляемый по обучающей выборке. Используется коэволюционный алгоритм безусловной однокритериальной оптимизации.

Питтсбургский этап формирования нечеткого классификатора. Индивиды представляют собой базу нечетких правил целиком. Длина хромосомы равна числу правил, найденных на Мичиганском этапе. Хромосомы бинарные, бит «1» означает использование соответствующего нечеткого правила, найденного на предыдущем этапе, бит «0» - исключение правила из базы. Пригодность – точность классификации базой правил. Вводится ограничение на максимально допустимое число правил, используемых в базе. Используется коэволюционный алгоритм условной однокритериальной оптимизации. Таким образом, Питтсбургский этап позволяет сформировать из полученных на предыдущем этапе правил базу, состоящую из существенно меньшего числа правил.

Для апробации гибридной схемы генерации нечеткого классификатора с использованием коэволюционного алгоритма были взяты следующие практические задачи классификации из репозитория UCI (для некоторых задач таких, как задача о выдаче банковского кредита, в репозитории используется упрощенная постановка с наличием только двух классов):

1. Credit (Australia-1) (задача о выдаче банковского кредита, австралийский вариант, 14 признаков, 2 класса);
2. Credit (Germany) (задача о выдаче банковского кредита, немецкий вариант, 24 признака, 2 класса);
3. Liver Disorder (диагностика заболевания печени, 6 признаков, 2 класса);
4. Iris (классификация видов ириса, 4 признака, 3 класса);
5. Yeast (классификация типов белков в дрожжах, 8 признаков, 10 классов);
6. Glass Identification (классификация сортов стекла по содержанию химических элементов, 9 признаков, 7 классов).

По данным таблицы 4 можно сделать выводы об особенностях работы каждого этапа гибридной схемы формирования нечеткого классификатора. Видно, что уже после формирования стартовых правил из обучающей выборки мы получаем вполне работоспособную базу. При этом число сгенерированных правил невелико по сравнению с общим количеством всех возможных правил.

Далее, Мичиганский этап позволяет несколько увеличить точность классификации базой нечетких правил. Таким образом, Мичиганский этап необходим, в первую очередь, для сглаживания эффектов случайности при формировании начальной базы нечетких правил.

Питтсбургский этап позволяет сократить число используемых нечетких правил. Компактные базы правил целесообразнее с точки зрения задачи обнаружения скрытых знаний. Кроме того, во многих случаях точность классификации после Питтсбургского этапа превосходит точность классификации, полученной на предыдущем этапе, несмотря на существенное сокращение числа правил. Это означает, что «плохие» правила оказывают негативное влияние на эффективность базы в целом, что подтверждает целесообразность сокращения числа используемых правил даже с точки зрения собственно задачи классификации (не говоря уже о том, что классификация по базе правил меньшего объема требует меньшего числа вычислений).

Таблица 4

Результаты исследования гибридной схемы генерирования нечеткого классификатора на задачах из репозитория UCI

Параметр	Задача					
	1	2	3	4	5	6
Начальное число правил	200	200	100	50	200	200
Средняя точность классификации начальной базы правил	0,854	0,782	0,595	0,945	0,417	0,640
Средняя точность классификации после Мичиганского этапа	0,870	0,791	0,653	0,979	0,495	0,687
Средняя точность классификации после Питтсбургского этапа (в скобках – ограничение на число используемых правил)	0,827(10) 0,861(20) 0,873(30)	0,762(50) 0,790(80)	0,666(10) 0,682(15) 0,692(30)	0,908(3) 0,951(4) 0,971(5) 0,975(6)	0,573(20) 0,586(30) 0,593(60)	0,737(20) 0,781(30)
Максимальная точность классификации после Питтсбургского этапа	0,870(10) 0,890(20) 0,891(30)	0,767(50) 0,794(80)	0,687(10) 0,710(15) 0,725(20)	0,947(3) 0,973(4) 0,987(5) 0,987(6)	0,598(20) 0,606(30) 0,626(60)	0,757(20) 0,827(30)

Разработанный метод генерации нечеткого классификатора обладает статистической устойчивостью – при проведении многократных запусков наблюдается незначительный разброс в значениях точности классификации полученных баз нечетких правил, а также генерация одних и тех же или семантически близких нечетких правил.

Из таблицы 5 видно, что нечеткий классификатор по точности решения задач превосходит многие современные алгоритмы классификации, что доказывает целесообразность использования данного алгоритма.

Сравнение различных алгоритмов классификации на задачах из репозитория UCI по точности классификации

Алгоритм	Credit (Australia-1)	Credit (Germany)	Liver Disorder
Нечеткий классификатор	0,891	0,794	0,725
Байесовский подход	0,847	0,679	0,629
Многослойный персептрон	0,833	0,716	0,693
Бустинг	0,760	0,700	0,656
Бэггинг	0,847	0,684	0,630
Метод случайных подпространств	0,852	0,677	0,632
Коэволюционный метод обучения алгоритмических композиций	0,866	0,746	0,644

В **четвёртой** главе приведено описание программных систем, реализующих разработанные алгоритмы.

Первая программная система «Коэволюционный алгоритм условной многокритериальной оптимизации» (разработана в среде C++ Builder 6.0) представляет собой универсальный комплекс для решения задач оптимизации различных классов и различного уровня сложности. Вторая программная система «Нечеткий коэволюционный классификатор» (разработана в среде MS Visual Studio 2005, язык C++) представляет собой универсальный комплекс для решения задач классификации с одновременным выявлением лингвистических знаний в соответствующих проблемных областях.

В четвертой главе также приведено детальное описание решения программными системами двух практических задач – задачи составления расписания посадок самолетов первой системой (как вычислительно сложной задачи условной оптимизации) и задачи распознавания спутниковых изображений второй системой (в рамках проекта «Студенческий спутник СибГАУ» в сотрудничестве с Центром управления полетами СибГАУ).

Составление расписания посадок самолетов в аэропорту состоит в установлении очередности приземления воздушных судов и назначении каждому из них времени приземления. При этом основными ограничительными критериями являются интервалы безопасности между посадками самолетов, обусловленные аэродинамическими свойствами. Критерий оптимальности – минимизация затрат, связанных с задержками посадок самолетов.

В задаче распознавания спутниковых изображений рассматривается шесть различных классов: красноземные почвы (1), хлопковые посева (2), сероземные почвы (3), влажные сероземные почвы (4), земли с овощными посевами (5), очень влажные сероземные почвы (6). Необходимо на снимке, полученном со спутника дистанционного зондирования Земли, отнести каждый квадрат размером 3×3 к тому или иному классу по значениям интенсивности в четырёх спектрах (зеленый, красный и два инфракрасных). Ниже приведена полученная при использовании нечеткого классификатора компактная база нечетких правил (обозначения лингвистических термов: ОМ – «очень малая интенсивность», М – «малая интенсивность», С – «средняя интенсивность», Б – «большая интенсивность», ОБ – «очень большая интенсивность»).

База нечетких правил для распознавания спутниковых изображений

№	Зеленый спектр	Красный спектр	1-й инфракрасный спектр	2-й инфракрасный спектр	Класс	Доверит. уровень
1	ОМ	С		М	1	0,899171
2				ОБ	2	1
3	ОМ	ОМ	С	С	2	0,921344
4		М	Б		2	0,898492
5	ОБ	ОБ			3	0,862182
6	Б		С	С	3	0,665146
7		Б		ОМ	4	0,592109
8	ОМ		ОМ		5	0,959055
9	М	М		М	5	0,757215
10		С		ОМ	6	0,675657

Точность классификации по базе правил на обучающей выборке – 0,8490, на тестовой – 0,8435.

Решенные практические задачи подтверждают обоснованность, достоверность и реализуемость предложенных методов и алгоритмов.

В **заключении** диссертации приведены основные результаты и выводы.

ОСНОВНЫЕ РЕЗУЛЬТАТЫ И ВЫВОДЫ

1. На основе анализа состояния проблемы построения систем классификации обоснована целесообразность разработки нечетких классификаторов и автоматизации их проектирования с помощью самонастраивающихся эволюционных алгоритмов глобальной оптимизации.

2. Разработан новый самонастраивающийся коэволюционный алгоритм условной оптимизации, отличающийся от известных самонастройкой метода учёта ограничений и модифицированным оператором миграции (пропорционально-групповая схема) и не уступающий им в эффективности решения задачи.

3. Разработан новый самонастраивающийся коэволюционный алгоритм многокритериальной оптимизации, отличающийся от известных отсутствием независимой работы индивидуальных алгоритмов в коэволюционной комбинации и способом оценивания индивидуальных алгоритмов.

4. Проведены комплексные исследования эффективности коэволюционного алгоритма как инструмента автоматизации выбора настроек генетического алгоритма на различных классах оптимизации. Наибольший эффект от использования коэволюционного подхода демонстрируется на задачах условной однокритериальной оптимизации, где коэволюционный алгоритм эффективнее стандартного генетического алгоритма с наилучшими настройками.

5. Обоснована целесообразность использования коэволюционного подхода в параллельных вычислениях, показано значительное повышение его эффективности при распараллеливании.

6. Разработан новый метод формирования нечеткого классификатора, отличающийся от известных новой схемой комбинирования Мичиганского и Питтсбургского подходов, наличием формализованной процедуры отбора

стартовых нечетких правил и использованием самонастраивающихся коэволюционных алгоритмов.

7. Проведены исследования предложенного метода формирования нечеткого классификатора на множестве практических задач классификации. По результатам исследований установлено, что данный метод позволяет получать эффективную базу правил при использовании небольшого числа стартовых правил и в итоге формировать компактные базы нечетких правил, обладает статистической устойчивостью и позволяет получать нечеткие классификаторы, превосходящие по эффективности многие современные алгоритмы классификации.

8. Разработанные в ходе исследования алгоритмы и методы реализованы в виде двух программных систем, успешно использованных затем при решении реальных практических задач.

Таким образом, в диссертации разработаны, исследованы и апробированы новые коэволюционные алгоритмы для автоматизации выбора настроек генетических алгоритмов и новый метод формирования нечеткого классификатора самонастраивающимися коэволюционными алгоритмами для решения задач классификации и интеллектуального анализа данных, что является существенным вкладом в теорию и практику исследования методов преобразования информации в данные и знания, методов работы со знаниями, методов машинного обучения и обнаружения новых знаний.

ПУБЛИКАЦИИ ПО ТЕМЕ ДИССЕРТАЦИИ

Статьи в ведущих рецензируемых научных журналах и изданиях

1. Сергиенко, Р.Б. Коэволюционный генетический алгоритм решения сложных задач условной оптимизации / Р.Б. Сергиенко, Е.С. Семенкин // Вестник Сибирского государственного аэрокосмического университета имени академика М.Ф. Решетнёва. – №2 (23). – 2009. – С. 17-21.

2. Сергиенко, Р.Б. Исследование эффективности коэволюционного генетического алгоритма условной оптимизации / Р.Б. Сергиенко // Вестник Сибирского государственного аэрокосмического университета имени академика М.Ф. Решетнёва. – №3 (24). – 2009. – С. 31-36.

3. Сергиенко, Р.Б. Метод формирования нечеткого классификатора самонастраивающимися коэволюционными алгоритмами / Р.Б. Сергиенко // Искусственный интеллект и принятие решений. – №3. – 2010. – С. 98-106.

4. Сергиенко, Р.Б. Коэволюционный алгоритм для задач условной и многокритериальной оптимизации / Р.Б. Сергиенко, Е.С. Семенкин // Программные продукты и системы. – №4. – 2010. – С. 141-144.

Публикации в журналах и сборниках

5. Сергиенко, Р.Б. Системы поддержки принятия решений на основе коэволюционного вероятностного алгоритма / Р.Б. Сергиенко, Е.С. Семенкин // Ползуновский альманах. - № 4. – 2008. – С. 217-219.

6. Сергиенко, Р.Б. Коэволюционный алгоритм условной оптимизации «Conditional Coevolution» / Р.Б. Сергиенко // Компьютерные учебные программы и инновации. – № 2. – 2009. – С. 21.

Публикации в сборниках трудов конференций

7. Sergienko, R. Competitive Cooperation for Strategy Adaptation in Coevolutionary Genetic Algorithm for Constrained Optimization / R. Sergienko, E.Semenkin // IEEE World Congress on Computational Intelligence (WCCI'2010), Barcelona, Spain, 2010. – P. 1626-1631.

8. Сергиенко, Р.Б. Разработка и исследование гибридной схемы формирования нечеткого классификатора с использованием коэволюционного алгоритма / Р.Б. Сергиенко // Труды Национальной конференции по искусственному интеллекту КИИ-2010. – Т.2. – М.: Физматлит, 2010. – С. 353-360.

9. Сергиенко, Р.Б. Гибридная схема генерирования нечеткого классификатора с использованием коэволюционного алгоритма / Р. Б. Сергиенко // Теория и практика системного анализа. Труды I Всероссийской научной конференции молодых ученых. – Т.1. – Рыбинск: РГАТА, 2010. – С. 11-19.

10. Сергиенко, Р.Б. Нечеткий коэволюционный классификатор / Р.Б. Сергиенко // Технологии Microsoft в теории и практике программирования: сборник трудов VII Всероссийской научно-практической конференции. – Томск: Издательство Томского политехнического университета, 2010. – С. 94-95.

11. Сергиенко, Р.Б. Нечеткий генетический классификатор в задаче распознавания спутниковых изображений / Р.Б. Сергиенко // Информационные технологии и математическое моделирование. Материалы VIII Всероссийской научно-практической конференции в 2-х частях. – Томск: Издательство Томского университета, 2009. – Часть 2. – С. 272-276.

12. Сергиенко, Р.Б. Параллелизация коэволюционного алгоритма в многопроцессорных вычислительных системах / Р.Б. Сергиенко // Решетневские чтения. Материалы XIII Международной научной конференции в 2-х частях. – Красноярск: СибГАУ, 2009. – Часть 2. – С. 453-454.

13. Сергиенко, Р.Б. Использование коэволюционного алгоритма при извлечении знаний в виде нечетких правил в задачах классификации / Р.Б. Сергиенко // Труды Конгресса по интеллектуальным системам и информационным технологиям "AIS-IT'09". Научное издание в 4-х томах. – М.: Физматлит, 2009. – Т.3. – С. 321-328.

14. Сергиенко, Р.Б. Составление расписания посадок самолетов коэволюционным алгоритмом условной оптимизации / Р.Б. Сергиенко // Решетневские чтения. Материалы XII Международной научной конференции. – Красноярск: СибГАУ, 2008. – С. 308-310.

15. Сергиенко, Р.Б. Коэволюционный алгоритм условной оптимизации: разработка и приложения / Р.Б. Сергиенко // Труды Международных научно-технических конференций "Интеллектуальные системы" (AIS'08) и "Интеллектуальные САПР" (CAD-2008). Научное издание в 4-х томах. – М.: Физматлит, 2008. – Т.1. – С. 33-40.

16. Сергиенко, Р.Б. Автоматизированное генерирование интеллектуальных информационных технологий на основе самонастраивающихся эволюционных алгоритмов / Р.Б. Сергиенко, В.В. Бухтояров // Технологии Microsoft в теории и практике программирования. – Новосибирск: НГУ, 2008. – С. 167-169.

17. Сергиенко, Р.Б. Автоматическая настройка параметров генетического алгоритма коэволюционным методом в решении задач условной оптимизации / Р.Б. Сергиенко // Молодежь и современные информационные технологии. Сборник трудов VI Всероссийской научно-практической конференции. – Томск: Издательство ТПУ, 2008. – С. 277-278.

18. Сергиенко, Р.Б. Коэволюционный алгоритм решения задач условной оптимизации / Р.Б. Сергиенко // VIII Всероссийская конференция молодых ученых по математическому моделированию и информационным технологиям. – Новосибирск: Институт вычислительных технологий СО РАН, 2007. – С.73.

19. Сергиенко, Р.Б. Разработка и исследование коэволюционного алгоритма для решения задач условной оптимизации / Р.Б. Сергиенко // Информационные технологии и математическое моделирование (ИТММ-2007): Материалы VI Международной научно-практической конференции. – Томск: Издательство Томского университета, 2007. – Ч.2. – С.96-99.

20. Сергиенко, Р.Б. Исследование эффективности коэволюционного алгоритма на тестовых функциях одной переменной / Р.Б. Сергиенко // VII Всероссийская конференция молодых ученых по математическому моделированию и информационным технологиям. – Красноярск: Институт вычислительного моделирования СО РАН, 2006. – С.68.

Зарегистрированные программные системы

21. Сергиенко, Р.Б. Коэволюционный алгоритм условной оптимизации «Conditional Coevolution»: свидетельство об отраслевой регистрации разработки. – М.: ВНИИЦ, 2008. – № гос. рег. 50200802249.

22. Сергиенко, Р.Б. Коэволюционный алгоритм условной многокритериальной оптимизации: свидетельство о государственной регистрации программы для ЭВМ / Р.Б. Сергиенко, Е.С. Семенкин. – М.: Реестр программ для ЭВМ, 2009. – № гос. рег. 2009616049.

23. Сергиенко, Р.Б. Нечеткий коэволюционный классификатор: свидетельство о государственной регистрации программы для ЭВМ / Р.Б. Сергиенко. – М.: Реестр программ для ЭВМ, 2010. – № гос. рег. 2010611766.

Сергиенко Роман Борисович

Автоматизированное формирование нечетких классификаторов
самонастраивающимися коэволюционными алгоритмами

Автореферат

Подписано к печати 27.10.2010. Формат 60x84/16

Уч. изд. л. 1.0 Тираж 100 экз. Заказ № _____

Отпечатано в отделе копировальной и множительной техники СибГАУ.

660014, г. Красноярск, пр. им. газ. «Красноярский рабочий», 31