На правах рукописи

Алхасов Станислав Сергеевич

РАЗРАБОТКА И ИССЛЕДОВАНИЕ ОПТИМИЗАЦИОННЫХ АЛГОРИТМОВ ДЛЯ РЕШЕНИЯ ЗАДАЧ БИНАРНОЙ КЛАССИФИКАЦИИ

Специальность 05.13.17 – Теоретические основы информатики

АВТОРЕФЕРАТ

диссертации на соискание ученой степени

кандидата технических наук

Таганрог – 2018

Работа выполнена в ФГАОУ ВО «Южный федеральный университет» на кафедре информационно-аналитических систем безопасности Института компьютерных технологий и информационной безопасности

Научный руководитель: доктор технических наук, профессор
Целых Александр Николаевич,
ФГАОУ ВО «Южный федеральный университет», заведующий кафедрой информационно-аналитических систем безопасности (г. Таганрог)

Официальные оппоненты: доктор технических наук, профессор
Карелин Владимир Петрович,
ЧОУ ВО «Таганрогский институт управления и экономики», заведующий кафедрой прикладной математики и информационных технологий

кандидат технических наук, доцент
Береза Андрей Николаевич,
Институт сферы обслуживания и предпринимательства (филиал) ФГБОУ ВО «Донской государственный технический университет», доцент кафедры «Информационные системы и радиотехника» (г. Шахты)

Ведущая организация: ФГБОУ ВО «Ростовский государственный университет путей сообщения»

Защита состоится « 28 » сентября 2018 г. в 11^{00} часов на заседании диссертационного совета Д 999.065.02 ФГАОУ ВО «Южный федеральный университет» по адресу: 347928, г. Ростовская область, г. Таганрог, пер. Некрасовский, 44, ауд. Д-406.

С диссертацией можно ознакомиться в Зональной научной библиотеке ФГАОУ ВО «Южный федеральный университет» по адресу: 344000, Ростовская область, г. Ростов-на-Дону, ул. Зорге, 21ж и на сайте: https://hub.sfedu.ru/diss/.

Отзыв на автореферат, заверенный гербовой печатью организации, просим направлять ученому секретарю диссертационного совета Д 999.065.02 по адресу: 347928, Ростовская область, г. Таганрог, пер. Некрасовский, 44, к. Г-144.

Автореферат разослан «___» _____ 2018 г.

Ученый секретарь
диссертационного совета Д 999.065.02
д-р техн. наук, профессор А.В. Боженюк

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Проблема нахождения эффективных решений задач бинарной классификации в условиях разнородности данных является основным мотивом выполнения настоящего диссертационного исследования.

Соискателем предлагается оптимизационный подход, позволяющий автоматизировать подбор параметров классификаторов и способный повысить качество выполнения классификации за конечный период времени.

Авторским вкладом в разработанном подходе являются предлагаемый исходный набор рассматриваемых методов классификации, новые критерии качества бинарной классификации, комбинированный оптимизационный алгоритм и в целом концепция автоматизации подбора параметров классификаторов.

Таким образом, выполненная соискателем работа посвящена разработке оптимизационных алгоритмов для повышения качества бинарной классификации с использованием оптимальных параметров, наиболее подходящих для того или иного метода классификации.

Актуальность. В настоящее время методы интеллектуального анализа данных получают широкое распространение в различных отраслях науки, техники и сферы услуг. Одна из важнейших групп методов – бинарная классификация – имеет ряд нерешенных проблем, среди которых высокоактуальна проблема эффективного автоматизированного выполнения классификации в условиях периодически изменяющейся структуры анализируемых данных, содержащих пропуски, выбросы и повторяющиеся идентичные записи. Важной классом задач бинарной классификации, рассматриваемых в настоящем исследовании, является проблема классификации объектов разбалансированных выборок, содержащих избыточную и неточную информацию. К этому классу задач относится задача удержания потребителей, представляющая собой одну из важнейших маркетинговых проблем для любого современного предприятия, работающего в высококонкурентных сферах телекоммуникаций, банкинга и страхования. На примере вышеуказанной прикладной проблемы анализа данных выполнены исследования в настоящей диссертационной работе.

К настоящему моменту существует ряд работ в области класса задач бинарной классификации объектов разнородных выборок. М.А.Х. Фаркад (Farquad), А. Родан (Rodan), Хуан Бинкуан (Huang Bingquan), Т. Вафеиадис (Vafeiadis), Хуан Йин (Huang Ying), Т. Кечади (Kechadi), А. Керамати (Keramati) и ряд других авторов выполнили большой объем работы по выявлению оптимальных реализаций всех этапов интеллектуального анализа данных от подготовки исходных данных до визуализации полученных результатов применительно к прогнозированию оттока потребителей и прочим задачам подобного типа. Вместе с тем специфика бинарной классификации в целом и данной прогностической задачи в частности такова, что каждый исследователь имеют свою собственную разновидность исследуемой задачи,

прежде всего базирующуюся на практически неограниченном разнообразии анализируемых исходных данных и на специфике доступных вычислительных ресурсов. Среди рассматриваемых вышеуказанными учеными методов бинарной классификации наибольшее распространение находят искусственные нейронные сети, деревья решений, метод опорных векторов и др.

В последнее время делаются попытки разработать более общие подходы, позволяющие расширить применимость существующих прогностических моделей. Так в исследованиях О.Е. Бухарова и Д.П. Боголюбова предложено использовать генетические алгоритмы для отбора наиболее информативных входных признаков, далее анализируемых искусственной нейронной сетью. Совместное рассмотрение искусственных нейронных сетей и генетических алгоритмов, вообще говоря, встречается во множестве работ, однако большинство из них сфокусировано на оптимизации функционала качества обучения нейросетей посредством генетических алгоритмов для решения специфических задач классификации и регрессии, когда традиционные методы оптимизации представляются менее предпочтительными. Среди таких работ следует отметить исследования В.А. Мищенко, А.А. Коробкина, Ши Хуавана (Shi Huawang), А.А. Олейника, С.А. Субботина, Ю.В. Чернухина, М.А. Беляева, Л.М.Л. де Кампоса (Campos) и др.

В работах В.М. Курейчика, В.В. Курейчика, Д. Уитли (Whitley), Ю.Р. Цоя, В.Г. Редько, Х.М. Пандей (Pandey) и др. показано, что генетические алгоритмы являются высокоэффективными, модифицируемыми и широко применимыми оптимизационными методами, моделирующими процесс биологической эволюции посредством операторов селекции, скрещивания (кроссинговера) и мутации. При этом они являются менее узкоспециализированными по сравнению со значительным числом традиционных методов оптимизации. Также в контексте рассматриваемой прогностической задачи важно то, что генетические алгоритмы не нуждаются в дифференцируемости целевой функции. Соответственно, применимость генетических алгоритмов не ограничивается их использованием для отбора входных признаков для нейронных сетей.

Исходя из вышесказанного, была предложена концепция оптимизации посредством генетических алгоритмов ряда разнородных параметров, характеризующих применяемые для прогнозирования классификаторы, для выявления такой архитектуры классификатора, которая обеспечивает наивысшую эффективность прогнозирования. К задаче оптимизации в рамках данного подхода относится отбор признаков, анализ эффективности того или иного способа нормализации данных и перемешивания объектов в выборке, определение числа блоков перекрестной проверки в процессе обучения классификатора и выявление оптимального набора параметров классификатора (например, числа ближайших соседей для одноименного метода, числа нейронов в скрытом слое для нейросети и т. д.).

Цель и задачи исследования. Целью исследования является разработка оптимизационного алгоритма для повышения точности бинарной класси-

фикации за счет определения оптимальных параметров того или иного метода бинарной классификации.

Для достижения поставленной цели требуется решить ряд задач:

1. Анализ эффективности известных методов интеллектуального анализа данных в контексте решения класса задач бинарной классификации.
2. Анализ и исследование оценок эффективности прогностической модели с учетом не только точности классификации, но и длительности работы классификатора.
3. Разработка генетического алгоритма, позволяющего выявлять оптимальные параметры классификатора, определяющие его точность, за ограниченный период времени.
4. Экспериментальное исследование разработанного комплексного оптимизационного подхода на примере оценки лояльности потребителей телекоммуникационных услуг.

Объект исследования. Объектом настоящего исследования является разнородная информация технического и коммерческого характера о потребителях услуг телекоммуникационного предприятия, содержащая пропущенные значения, выбросы (outliers), повторяющиеся идентичные записи и коррелированные между собой признаки.

Предмет исследования. Предметом исследования в данной диссертационной работе является бинарная классификация объектов разнородных выборок с изменчивой структурой данных и неравнозначными классами, оптимизируемая генетическими алгоритмами.

Методы исследования. В диссертационной работе использованы методы интеллектуального анализа данных, такие как искусственные нейронные сети, метод k ближайших соседей, деревья решений и метод опорных векторов для решения задач классификации и генетические алгоритмы с целью оптимизации.

Научная новизна работы. Научная новизна работы состоит в следующем:

1. Разработан комбинированный генетический алгоритм, позволяющий находить оптимальный набор параметров алгоритмов бинарной классификации, отличающийся от известных сочетанием катастрофической и островной моделей со специализацией островов.
2. Разработаны и исследованы критерии качества бинарной классификации, позволяющие учитывать разбалансированность классифицируемых объектов выборок и длительность выполнения бинарной классификации, отличающиеся от известных возможностью однозначной трактовки числовых значений критериев качества.
3. Предложена универсальная методика автоматизации подбора параметров алгоритмов бинарной классификации, учитывающая возможность выполнения бинарной классификации в автоматизированном режиме без постоянного экспертного контроля, отличающаяся применением разработанного генети-

ческого алгоритма и критериев качества бинарной классификации для однозначной оценки эффективности выполнения классификации.
4. Разработан и внедрен программный комплекс, осуществляющий бинарную классификацию объектов разбалансированных выборок, отличающийся использованием алгоритмов оптимизации параметров классификаторов, что позволяет получать для каждой конкретной ситуации такой классификатор, который обеспечивает наивысший уровень качества бинарной классификации.
5. На основе разработанного программного комплекса проведен анализ нового класса задач бинарной классификации, отличающихся разнородностью анализируемых выборок, неравнозначностью классов и изменчивостью структуры данных.

Практическая значимость. Практически значимыми являются разработанные критерии качества бинарной классификации, новая комбинированная реализация генетического алгоритма на основе островной модели и модели эволюции катастроф Г. де Фриза для решения оптимизационных задач и обобщенные сведения о применении методов бинарной классификации в решении прикладных задач на примере прогнозирования оттока потребителей.

Реализация и внедрения результатов работы. Описанная в настоящей работе концепция реализована в программном продукте «Система классификаторов для прогнозирования оттока потребителей услуг телекоммуникационного предприятия», для которого получено Свидетельство о государственной регистрации программы для ЭВМ №2016662656. Дата государственной регистрации в Реестре программ для ЭВМ – 17 ноября 2016 г.

Основные результаты и положения диссертационной работы внедрены в учебном процессе Южного федерального университета на кафедре информационно-аналитических систем безопасности Института компьютерных технологий и информационной безопасности (г. Таганрог), а также применены в деятельности ООО «Южные телефонные сети» (г. Ростов-на-Дону) и ООО «Интеллектика Консалтинг» (г. Ростов-на-Дону).

Апробация работы. Основные положения и результаты работы диссертационной работы докладывались и обсуждались на российских и международных научно-технических конференциях: Всероссийской научной конференции «Системы и модели в информационную эпоху» (г. Таганрог, апрель 2014 г., СМИЭ-2014); VIII Международной научной конференции «Security of Information and Networks» (г. Сочи, 8 – 10 сентября 2015 г., SIN 2015); XXIII Научной конференции «Современные информационные технологии: тенденции и перспективы развития» (г. Ростов-на-Дону, 21 – 22 апреля 2016 г., СИТО-2016); III Международной научной конференции «Information Technologies in Science, Management, Social Sphere and Medicine» (г. Томск, май 2016 г., ITSMSSM 2016); IV Международной научной конференции «Information Technologies in Science, Management, Social Sphere and Medicine» (г. Томск, декабрь 2017 г., ITSMSSM 2017).

Публикации. По теме диссертации опубликовано 10 работ, из них 5 статей в изданиях, входящих в перечни ВАК РФ, Scopus и Web of Science. Все результаты, составляющие основное содержание диссертации, и выносимые на защиту положения получены и сформулированы диссертантом самостоятельно. Большинство работ [1–4, 6–10] опубликовано в соавторстве с научным руководителем А.Н. Целых, которому принадлежит постановка задачи и разработка концепции исследования. Часть работ имеет соавтора А.А. Целых [2–5, 9], у которого диссертант получал многочисленные консультации по интеллектуальному анализу данных.

Структура и объем работы. Диссертация состоит из введения, трех глав, заключения, списка литературы из 99 наименований и двух приложений. Основное содержание диссертации включает текст, 37 рисунков и 24 таблицы общим объемом 137 страниц. Полный объем диссертационной работы составляет 156 страниц.

Область исследования. Диссертационная работа соответствует пунктам 5 и 13 паспорта научной специальности 05.13.17 – Теоретические основы информатики.

П.5. Разработка и исследование моделей и алгоритмов анализа данных, обнаружения закономерностей в данных и их извлечениях, разработка и исследование методов и алгоритмов анализа текста, устной речи и изображений.

П.13. Применение бионических принципов, методов и моделей в информационных технологиях.

Основные положения, выносимые на защиту. На защиту выносятся:

1. Разработанные новые критерии качества бинарной классификации – взвешенная полнота, оценка взвешенной полноты и длительности.
2. Метод автоматизированного подбора параметров алгоритмов бинарной классификации и анализируемых данных для повышения эффективности прогнозирования.
3. Разработанный соискателем генетический алгоритм, комбинирующий островную модель со специализацией каждого острова и модель эволюции катастроф Г. де Фриза.

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

Во **введении** обосновывается актуальность темы, приведены новизна и практическая значимость исследования, формулируются основные цели и задачи, представлены основные положения, выносимые на защиту данной диссертационной работы.

В **первой главе** приведен обзор задач бинарной классификации на примере прогнозирования оттока (churn prediction) потребителей высокотехнологичных услуг (телекоммуникационных, банковских, услуг страхования и др.) в условиях современного высококонкурентного рынка и описаны особенности оценки эффективности ее решения. Показано, что в случае оценки

лояльности потребителей прогнозирование их оттока сводится к задаче бинарной классификации (класс –1 – нелояльные, класс 1 – лояльные потребители). Представлен обзор работ по прогнозированию оттока. Проанализированы важнейшие методы классификации и выделены основные из них, наиболее эффективные в рамках рассматриваемой задачи. Из множества разнообразных стандартных характеристик качества классификации выделен критерий, наилучшим образом соответствующий рассматриваемой прогностической задаче. Этот критерий называется полнотой (recall) и задается формулой

$$\mu = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \qquad (1)$$

где TP – число истинно положительных распознаваний,
FN – число ложноотрицательных распознаваний.

Среди разных способов классификации, относящихся к методам интеллектуального анализа данных (data mining), на основании рассмотренного массива работ выбраны дерево решений C4.5, метод k ближайших соседей (kNN), метод опорных векторов (SVM) и искусственные нейронные сети (ИНС). Сделать теоретические заключение о том, что некоторый алгоритм бинарной классификации является оптимальным в общем случае нельзя, потому что согласно NFL-теореме (Д. Вольперт, У. Макреди) универсальных алгоритмов, решающих разнообразные задачи с одинаково высокой точностью, не существует. Помимо этого, можно также отметить, что структура собираемых и анализируемых данных с течением времени может существенно меняться по разным причинам технического и коммерческого характера, вследствие чего ранее отобранный в ручном режиме алгоритм классификации может оказаться гораздо менее эффективным. Это может случиться, например, когда увеличивается число признаков (features), появляются номинативные и псевдовещественные (напр., почтовый индекс) признаки и/или когда появляется большое число дублей в собираемых данных или часть вновь полученных объектов (записей, samples) содержит пропущенные значения и выбросы.

Вторая глава посвящена исследованию эффективности методов интеллектуального анализа данных в задачах прогнозирования оттока потребителей на примере телекоммуникационной отрасли. Изучена прогностическая способность четырех выбранных методов классификации на основании ряда общих рекомендаций. По завершению проведенных исследований среди разных вариаций решающего дерева C4.5 было выбрано дерево без ограничений на глубину, число листовых вершин и число объектов в таких вершинах. Среди всех возможных вариантов метода k ближайших соседей был отобран kNN с $k = 3$ и Манхэттенской мерой расстояния. В машине опорных векторов была использована полиномиальная ядерная функция степени $p = 3$ с коэффициентом $\gamma = 0$ и константой регуляризации $C = 1000$. Искусственная нейронная сеть была опытным путем сформирована из двух скрытых слоев, содержащих 25 и 15 нейронов. При этом нейросеть была дополнена коэффи-

циентом импульса (momentum), равным 0,8. Было показано, что некоторые широко известные по ранним работам в области ИНС подходы по подбору оптимального числа нейронов на практике в условиях реальных разнородных данных могут оказываться ограниченно применимыми.

Нейросетевой классификатор оказался наиболее эффективным для решения задачи прогнозирования оттока. Однако, вместе с тем, многочисленные эксперименты с разнообразными реализациями бинарных классификаторов показали, что применяемые критерии качества классификации не всегда адекватно отражают прогностическую способность классификатора. Иными словами, высокие значения ранее упомянутой основного критерия в прогнозировании оттока – *полноты* (recall) – не всегда свидетельствуют о высоком качестве бинарной классификации. Всякий раз исследователю требуется изучать значения вспомогательных критериев, таких как *доля верных распознаваний* (accuracy) и *точность* (precision). Таким образом, можно выделить две проблемы, важность которых многократно возрастает, когда проводится большое число запусков различных классифицирующих алгоритмов:

1. Время выполнения алгоритма классификации: Некоторые алгоритмы, например, машина опорных векторов выполняются существенно дольше других классификаторов, но при этом получаемые результаты оказываются ниже или же незначительно выше по сравнению с другими, более быстрыми классифицирующими алгоритмами.
2. Необходимость учитывать ложноположительные распознавания (класс 1 вместо класса –1) и соотношение между классами: В противном случае при автоматизации рутинных операций подбора оптимального для текущей задачи классификатора в отсутствие постоянного экспертного контроля всех основных критериев качества для каждого рассматриваемого алгоритма классификации подбор классификатора сведется, в конечном счете, к получению в качестве оптимального результата вырожденного константного классифицирующего алгоритма, положительно (класс 1) распознающего все объекты выборки. Иными словами, в случае такого псевдооптимального классификатора, характеризующегося наивысшими значениями полноты, все потребители будут охарактеризованы как склонные к оттоку, что не имеет никакой практической ценности.

С целью преодоления вышеобозначенных проблем разработаны два новых критерия качества бинарной классификации:

1. Взвешенная полнота, которая учитывает число ложноположительных распознаваний и соотношение между двумя классами в обучающей выборке и определяется по формуле

$$\mu' = \mu\left(1-\left(\frac{\mathrm{FP}}{\mathrm{FP}+\mathrm{TP}}\right)^{10\delta}\right), \qquad (2)$$

где FP – число ложноположительных распознаваний,
TN – число истинно отрицательных распознаваний,

$$\delta = \frac{\mathrm{TP}+\mathrm{FN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}}.$$

2. Оценка взвешенной полноты и длительности (ОВПД), определяющая качество классификатора не только по величине взвешенной полноты, но и с учетом времени выполнения алгоритма бинарной классификации (рис. 1).

$$\omega' = \exp\frac{\mu' - \mu_0}{\max(t, t_m)} + \mu'^2 - \left(\frac{t}{t_0}\right)^2, \qquad (3)$$

где t – длительность выполнения алгоритма (в секундах),

μ_0 – пороговое значение полноты (принято на уровне 0,8),

t_0 – предельно допускаемое время выполнения алгоритма (принято на уровне 100),

t_m – корректировка минимально допустимого времени выполнения (принято на уровне 1).

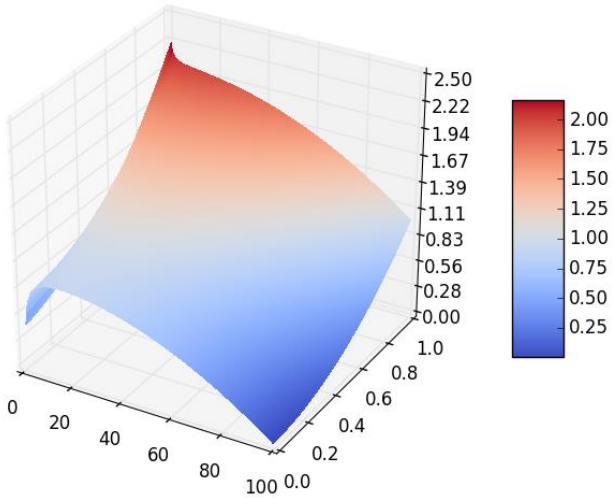

Рис. 1. Зависимость ОВПД от взвешенной полноты и времени выполнения алгоритма

Взвешенная полнота использована для анализа отдельных классификаторов в процессе их разработки и отладки, тогда как ОВПД применяется в тех случаях, когда в автоматизированном режиме подбирается наиболее эффективный для анализируемых данных классификатор. В табл. 1 сведены классифицирующие алгоритмы, определенные как оптимальные по критерию полноты. При этом для подтверждения недостаточной информативности этой стандартного критерия качества классификации приведены значения вновь разработанных критериев.

Из табл. 1 следует, что в одних случаях полнота и взвешенная полнота отличаются друг от друга значительно сильнее, чем в других случаях. Это можно показать на примере классификатора, основанного на методе опорных векторов, имеющего полиномиальное ядро 3-ей степени с $\gamma = 0$. Если обратиться к выражению полиномиальной ядерной функции, можно получить в текущем случае следующую запись

$$K(x_i, x) = (\gamma x_i x + c_0)^d = c_0^d, \qquad (4)$$

откуда следует, что тот факт, что любые верные распознавания не будут иметь какого-либо осмысленного обоснования.

Другой пример, следующий из табл. 1, – метод k ближайших соседей (kNN), имеющий максимальную полноту при k = 1, при том, что реализация 1NN хорошо известна своей неустойчивостью к шумам.

В процессе тщательного рассмотрения анализируемой выборки выявлен ряд особенностей, негативно влияющих на получаемые результаты классификации. Среди прочего обнаружен псевдовещественный признак «Почтовый индекс», в реальности имеющий скорее номинативный характер, нежели вещественный. Поэтому данный признак заменен совокупностью N фиктивных переменных (dummy variables), где N – число значений, которые может принимать признак «Почтовый индекс». Более того, фактически следует использовать $N + 1$ фиктивную переменную в связи с наличием в исходной выборке выбросов (т.е. когда в почтовом индексе отсутствует один (и более) десятичный разряд или же присутствует лишний разряд) и пропущенных значений, связанных с «человеческим фактором», обусловленным несовершенством процедуры сбора данных.

Табл. 1. Сравнение методов бинарной классификации по значениям полноты и модифицированных критериев качества

Метод		Решающее дерево C4.5	Метод ближайших соседей		Метод опорных векторов	Искусственные нейронные сети
Параметры метода		Без ограничений на глубину дерева, число объектов в листе и число листов	Манхэттенская мера расстояния		Полиномиальное ядро. $p = 3$, $C = 10^3$, $\gamma = 0$	25 и 15 нейронов в скрытых слоях. $\alpha = 0{,}8$
			$k = 1$	$k = 3$		
Критерии	μ	70,77%	45,29%	39,90%	85,09%	72,69%
	μ'	55,29%	28,98%	31,79%	19,94%	63,85%
	ω'	2,12	0,80	0,90	1,15	4,26

Табл. 2. Значения взвешенной полноты для классификаторов, анализирующих первоначальную выборку и выборку, модифицированную посредством алгоритма Add-Del

Метод / Выборка	Решающее дерево C4.5	kNN: $k = 3$	SVM-RBF: $C = 55$, $\gamma = 150$	ИНС: $\alpha = 0{,}8$
Первоначальная	55,29%	31,79%	46,05%	63,85%
Модифицированная по алгоритму Add-Del	58,68%	38,17%	41,48%	68,72%

С целью получения более качественной выборки, предобработанной в автоматизированном режиме без строго экспертного контроля, была проведена процедура отбора признаков (feature selection). Простейший способ отбора признаков – полный перебор (full search), на практике не реализуемый ввиду высокой ресурсозатратности. В качестве альтернативы полному перебору обычно предлагаются алгоритмы последовательного добавления признаков (Add), последовательного удаления признаков (Del) и Add-Del. По-

следний вариант является наиболее подходящим для задач с большим числом разнородных признаков, хотя и уступает по быстродействию алгоритму Add. В табл. 2 представлен результат классификации, получаемый в случае предобработки анализируемой выборки по алгоритму Add-Del.

Также было исследовано влияние перемешивания объектов (записей) выборки на качество получаемых прогностических моделей. Причиной этого рассмотрения явился тот факт, что в некоторых задачах интеллектуального анализа данных исследуемую выборку перемешивать не рекомендуется. Однако в общем случае считается желательным на этапе предварительной обработки данных случайным образом объекты выборки перемешивать. В рассматриваемой задаче прогнозирования оттока потребителей не было обнаружено преимуществ сохранения первоначального порядка объектов в данных, потому выборка была перемешана.

В **третьей главе** описана последовательность исследований, целью которых является разработка прогностической модели, в которой автоматизируется подбор алгоритма, наилучшим образом классифицирующего потребителей на лояльных (класс 1) и нелояльных (класс –1). Такого рода задача является оптимизационной. Целевой функцией здесь является критерий качества бинарной классификации, а именно ОВПД, чтобы исключить из рассмотрения классификаторы с низким быстродействием.

В главе 2 кратко были рассмотрены подходы к настройке нескольких основных классифицирующих алгоритмов с учетом различных теоретических и эмпирических знаний, а также с ограниченным использованием полного перебора. Очевидными недостатками этой совокупности подходов помимо больших затрат времени и вычислительных ресурсов являются, во-первых, опасность останова поиска наилучших классификаторов в локальных экстремумах и, во-вторых, необходимость постоянного экспертного контроля для интерпретации промежуточных результатов. Для преодоления этих недостатков в главе 3 предложено применение генетических алгоритмов для решения данной оптимизационной задачи.

Генетические алгоритмы представляют собой особый тип методов оптимизации. Важным их преимуществом является отсутствие существенных математических требований к виду целевой функции, называемой в случае эволюционных вычислений функцией приспособленности (fitness function), и ее ограничений. В генетическом алгоритме используются только значения самой целевой функции, а не ее производных. Еще одним преимуществом генетических алгоритмов является то, что они выполняют поиск оптимума не из одной точки, а одновременно из нескольких точек, называемых особями (individuals), образующими популяцию.

Разработка наиболее пригодного для рассматриваемой прогностической задачи генетического алгоритма велась последовательно в порядке усложнения его структуры. Первоначально был рассмотрен т.н. простой, или стандартный генетический алгоритм (ПГА), дополненный оператором эли-

тизма, предполагающим сохранение особи, соответствующей наилучшему значению функции приспособленности, из поколения в поколение.

В целях первоначальной оценки работоспособности генетический алгоритм был испытан на нескольких тестовых функциях (benchmark functions): сферической, Растригина, Розенброка (Rosenbrock) и Экли (Ackley). Полученные результаты свидетельствуют в целом о применимости данного подхода.

При этом качество оптимизации определялось достижением оптимума не просто за наименьшее число итераций (поколений популяции), а, прежде всего, за минимальное количество промежуточных вычислений функции приспособленности. Это связано с тем, что в рамках данного исследования требуется многократное выполнение процедуры бинарной классификации, которая в отличие от типичной тестовой функции принимает на вход не только набор параметров особи, а также еще выборку, которую требуется классифицировать и вычислить критерий качества бинарной классификации.

За ограниченное число (500000) вычислений функции приспособленности ПГА смог приблизиться лишь к глобальному минимуму сферической функции и функции Экли, тогда как в случае функций Растригина и Розенброка на выходе данного алгоритма оказались субоптимальные решения. Объяснением этому служит тот факт, что в случае ПГА имеет место тенденция к снижению разнообразия особей в ряду поколений, что приводит к попаданию конечного результата в один из локальных минимумов.

Для повышения разнообразия особей может быть предложено разделение их популяции на некоторое число изолированных ареалов, ограниченно связанных между собой за счет миграции особей раз в некоторое число поколений. Такая архитектура генетического алгоритма называется островной. Простой островной генетический алгоритм (ПОГА) оказывается более эффективным, чем ПГА за конечное число вычислений функции приспособленности (500000). Только лишь в случае функции Растригина он уступает ПГА. Важно отметить, что любой ПОГА характеризуется в частности числом островов и частотой миграций. В эксперименте на 500000 вычислений функции приспособленности приходилось 5 островов с 40 особями на каждом и 4 миграции. Такое разделение особей не обязательно является наиболее желательным, поэтому далее были рассмотрены вариации такого же ПОГА с разным числом островов и особей на каждом острове.

Из рис. 2 следует, что с ростом числа островов (при пропорциональном уменьшении числа особей на каждом острове) продуктивность ПОГА возрастает. В свою очередь сходимость алгоритма также улучшается. Таким образом, для оптимизации параметров алгоритмов бинарной классификации предлагается ПОГА с 25 островами.

В наиболее часто применяемой к настоящему времени в задачах машинного обучения библиотеке Scikit Learn для языка программирования Python подбор оптимального классификатора выполняется с помощью функции GridSearchCV, предусматривающей полный перебор всех заданных сочета-

ний параметров рассматриваемого классификатора. Однако, если требуется проанализировать значительное количество разных классификаторов за конечное время, данная функция оказывается неприменима. Более того, она не предусматривает в рамках единой вычислительной процедуры рассмотрение бинарных классификаторов разных типов.

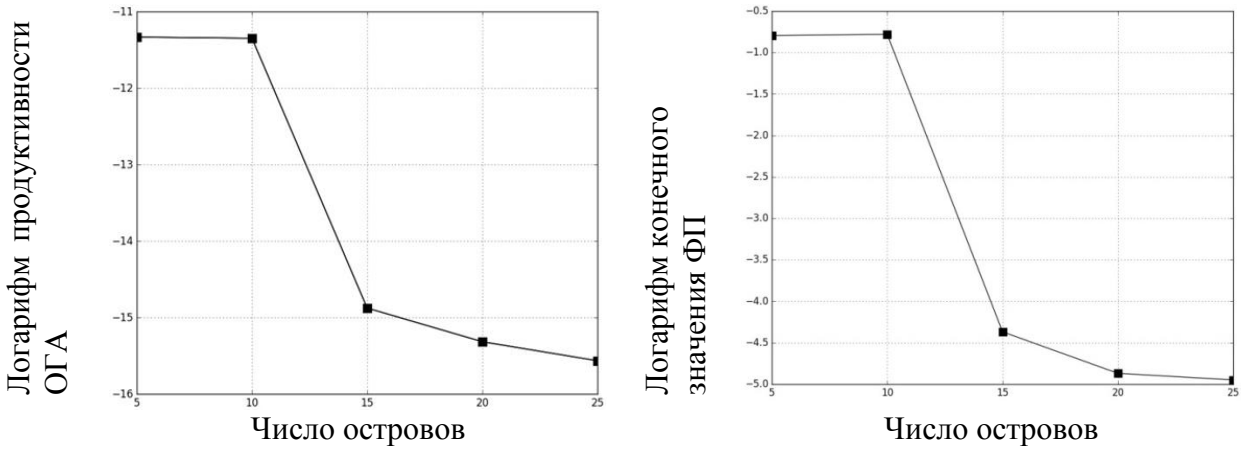

Рис. 2. Изменение десятичных логрифмов показателя продуктивности островного генетического алгоритма и конечного достигаемого значения функции приспособленности в зависимости от числа островов

Применение ПОГА позволило уменьшить число запусков классификаторов по сравнению с их полным перебором и при этом получить осмысленный результат. Для оценки конкурентоспособности особей на промежуточных итерациях ПОГА в качестве критерия качества бинарной классификации применялась ранее описанная оценка взвешенной полноты и длительности (ОВПД), тогда как для общей оценки генетического алгоритма применялась взвешенная полнота. Таким же образом оценивались и все последующие алгоритмы, предполагающие оптимизацию параметров бинарных классификаторов.

Полученные в случае использования ПОГА результаты отдельно по каждому типу классификатора были сравнены с наилучшими результатами, полученными посредством ограниченного перебора без применения оптимизации параметров классификаторов. Из табл. 3 видно, что применение ПОГА позволяет повысить качество бинарной классификации, но лишь незначительно (на 1,61%). Однако такой результат в условиях периодического изменения структуры анализируемых данных может оказаться даже ниже тех показателей, которые достигаются в случае ограниченного перебора. Поэтому ранее рассмотренный ПОГА следует модифицировать.

Одной из причин недостаточно высоких результатов оптимизации оказывается постепенная гомогенизация особей, в результате чего острова становятся дубликатами друг друга. Решить эту проблему можно посредством введения в генетический алгоритм оператора катастроф Г. де Фриза. Этот подход предполагает тот факт, что движущей силой эволюции является череда катастроф, некоторым образом удаляющих часть генофонда. В рассматри-

ваемой реализации генетических алгоритмов данная концепция предполагает, что при появлении тенденции к снижению дисперсий значений генов особей и соответствующих значений функции приспособленности через некоторое число поколений происходит катастрофа: произвольная часть особей на каждом острове удаляется (при этом действие оператора элитизма сохраняется), а удаленные особи заменяются произвольно инициализированными. Ниже данное описание приводится в строках 6–8 псевдокода (табл. 4).

Из табл. 3 следует, что островной алгоритм с катастрофизмом эффективнее ПОГА. Он по своей эффективности значительно (на 8,84%) превосходит лучшие результаты, получаемые посредством ограниченного использования полного перебора и учета известных эмпирических закономерностей. Однако его недостатком является большое число поколений, необходимых для достижения оптимума. Для преодоления данного недостатка предлагается в рамках концепции «исследования и использования» (exploration and exploitation) задавать разные значения вероятностей кроссинговера и мутации произвольным образом для каждого острова.

Табл. 3. Результаты применения оптимизационных алгоритмов для различных методов бинарной классификации в задаче прогнозирования оттока потребителей с использованием взвешенной полноты в качестве итогового критерия качества классификации

Методы классификации / Оптимизационные алгоритмы	Дерево решений	Метод k ближайших соседей	Метод опорных векторов	Искусственные нейронные сети
Наилучшие результаты, полученные без применения оптимизации параметров классификаторов	58,68%	38,17%	41,48%	68,70%
Простой островной ГА	61,18%	46,22%	40,83%	70,31%
Островной ГА с катастрофизмом	63,94%	49,60%	42,18%	77,54%
Разработанный автором островной ГА с катастрофизмом, дополненный специализацией генетических операторов на каждом острове	**66,05%**	**53,82%**	**58,61%**	**83,12%**

Стратегии «исследования и использования» объясняют способность ГА точно находить «дно» экстремума (локального или глобального) и определять среди большого числа экстремумов глобальный. «Исследование» базируется на операторе мутации и позволяет находить ранее неизвестные участки пространства поиска. «Использование» основывается на операторе кроссинговера и позволяет улучшить ранее полученные результаты. Таким образом, в процессе мутации проще преодолевать локальные экстремумы, тогда как кроссинговер помогает достичь «дна» экстремума с высокой точностью.

Полученные результаты свидетельствуют о том, что островной ГА со случайными вероятностями кроссинговера и мутации, в котором каждый остров получает свою собственную специализацию, позволяет достичь более высоких значений качества классификации (на 14,42% эффективнее по сравнению с наилучшими результатами, полученными без оптимизации) и при этом сходится за меньшее число итераций. Таким образом, удается избежать избыточных запусков бинарных классификаторов.

Из табл. 4 видно, что разработанный автором генетический алгоритм позволяет достичь существенного увеличения качества классификации, заметного по возросшему значению взвешенной полноты. Выполненный вычислительный эксперимент свидетельствует о том, что в рамках текущих входных данных, первоначально содержавших пропуски, дубли и выбросы, оптимальным оказывается нейросетевой классификатор. Классификатор, основанный на методе опорных векторов, также при использовании разработанного соискателем комбинированного генетического алгоритма существенно оптимизируется, однако, все же уступает ИНС-классификатору. Что касается, остальных классификаторов (дерево C4.5 и kNN), то повысить их классифицирующую способность не удается.

Разработанный алгоритм, сочетающий модель эволюции катастроф де Фриза и островную модель со специализацией генетических операторов островов, на основании проведенных исследований может быть признан существенно более эффективным для использования в прогностических моделях в качестве инструмента подбора оптимальных параметров алгоритмов бинарной классификации. Описанный подход превосходит известные прежде способы, основанные как на эмпирическом подборе основных параметров классификаторов, так и на использовании ранее известных моделей эволюционных вычислений. В табл. 3 приведена сравнительная оценка стандартных подходов подбора оптимальных параметров классификаторов и разработанного генетического алгоритма.

Таким образом, разработанный генетический алгоритм позволяет достичь оптимального решения, превосходящего по своему качеству полученные в главе 2 результаты. Полученный набор значений отражает перечень параметров классификатора и форму представления и предварительной обработки анализируемых входных данных. Для рассматриваемого входного массива был получен следующий набор значений параметров, характеризующих модель прогнозирования оттока потребителей:

а). Подготовка данных и внешние факторы классификации: ИНС в качестве метода классификации; minmax-нормализация данных; перемешивание объектов обучающей выборки; 6 блоков кросс-валидации.

б). Набор используемых параметров для классификации: Число голосовых сообщений; Число минут звонков в послеполуденное время; Число звонков в службу поддержки пользователей; Тарифный план для международных звонков; Число минут звонков в вечернее время; Стоимость звонков в послеполу-

денное время; Стоимость звонков в вечернее время; Число международных звонков; Стоимость международных звонков.
в). Параметры используемого метода классификации (ИНС): 28 и 16 нейронов в 1-ом и 2-ом скрытых слоях соответственно; импульс: 0,79; скорость обучения: 0,002; функция активации: гиперболический тангенс.

Табл. 4. Псевдокод разработанного автором островного генетического алгоритма, дополненного специализацией генетических операторов на каждом острове

Вход: $\{P_i\}$ – первоначально инициализированные островные подпопуляции ($i = \overline{1, I}$, I – число островов), G – число поколений, m – периодичность миграций между островами, φ – допустимый предел дисперсии значений функции принадлежности, γ – допустимый предел среднего значения дисперсий генов особей, $\left\{p_i^{(cr)}\right\}$, $\left\{p_i^{(mut)}\right\}$ – вероятности кроссинговера и мутации на каждом острове.
1: **для всех** $g = \overline{1, G}$: 2: **если** $g \bmod m = 0$, **то**: 3: $Q := \text{shuffle}\{P_i\}$ – объединение и смешение особей всех островов 4: $\{P_i\} := \underset{I}{\text{split}}\, Q$ – разделение общего пула особей на I островных подпопуляций 5: **для всех** $i = \overline{1, I}$: 6: **если** $g > 1$ **И** $\varphi > \text{var}\{F_i\}$ **И** $\gamma > \text{avg var}\{P_{i,\kappa}\}$, **то**: ($\kappa = \overline{1, K}$, K – число генов в особи) 7: $r_i := \text{random}[0; 0{,}5]$ – случайная доля замещаемых особей на i-ом острове в g-ом поколении 8: $P_i := \text{disaster}(P_i, r_i)$ – удаление некоторой доли r_i особей подпопуляции и их замена вновь инициализированными особями 9: $P_i, F_i := \text{GO}\left(P_i, p_i^{(cr)}, p_i^{(mut)}\right)$ – выполнение селекции, кроссинговера, мутации и элитизма на i-ом острове в g-ом поколении 10: $f_{\min}^{(i)} := \min F_i$ – минимальное значение функции приспособленности на i-ом острове в g-ом поколении 11: $f_{\min} := \min\left\{f_{\min}^{(i)}\right\}$ – минимальное значение функции приспособленности на всех островах в g-ом поколении
Вернуть $f_{\min}$

ЗАКЛЮЧЕНИЕ

Настоящая диссертационная работа посвящена разработке оптимизационного алгоритма, позволяющего подбирать оптимальный набор параметров заданных алгоритмов классификации, для формирования классификатора, способного в конкретных условиях наиболее эффективно выполнять бинарную классификацию.

В ходе выполнения данной работы получены следующие основные результаты и сделаны определенные выводы:

1. Сформирован набор основных методов бинарной классификации для решения задач, в которых анализируемые выборки являются разбалансированными, а классы – неравноценными на примере прогнозирования оттока потребителей. Показано, что базовым стандартным критерием качества бинарной классификации в данном случае является полнота.
2. Разработаны новые критерии качества бинарной классификации – взвешенная полнота и оценка взвешенной полноты и длительности (ОВПД). Первая (статическая) из них служит в качестве оценки влияния числа ложных распознаваний минорного класса (нелояльных потребителей в случае прогнозирования оттока) и общей разбалансированности анализируемой выборки. Вторая (динамическая) учитывает время выполнения вычислений с использованием того или иного классификатора.
3. Предложена стратегия автоматизации подбора основных параметров используемых классификаторов, а также характеристик анализируемой выборки (тип нормализации, перемешивание объектов выборки и т.д.) с помощью генетических алгоритмов.
4. Разработан комбинированный генетический алгоритм, сочетающий катастрофическую и островную модели со специализацией генетических операторов. Данный алгоритм позволяет формировать такие вариации классификаторов, которые позволяют получить результаты бинарной классификации, заметно превосходящие все остальные результаты ранее использованных классификаторов, характеристики которых основаны на эмпирически выявленных закономерностях или на более простых реализациях генетических алгоритмов.
5. В качестве результата использования разработанного генетического алгоритма, построенного на основе островной модели со специализацией островов и модели эволюции катастроф, получен набор параметров классификатора и анализируемых данных, соответствующих наиболее высокому качеству прогнозирования. Таким образом, архитектура классификатора в случае анализа лояльности потребителей телекоммуникационных услуг оказывается следующей: искусственная нейронная сеть обратного распространения ошибки с 28 и 16 нейронами в скрытых слоях, гиперболическим тангенсом в качестве активационной функции, скоростью обучения и импульсом, равными 0,79 и 0,002 соответственно. Данные при этом нормализуются по формуле minmax-нормализации и перемешиваются. Число блоков перекрестной про-

верки (кросс-валидации) устанавливается равным шести. Также выявляется оптимальное сочетание анализируемых признаков, количество которых ограничивается девятью.

ПУБЛИКАЦИИ ПО ТЕМЕ ДИССЕРТАЦИИ

Статьи, опубликованные в изданиях, включенных в перечень ВАК РФ или индексируемых научными базами данных Scopus и Web of Science

1. Алхасов С.С., Целых А.Н. Основные подходы к построению информационной системы для моделирования оттока клиентов услуг связи // Известия ЮФУ. Технические науки. – № 2 (163). – 2015. – С. 106–115.
2. Alkhasov S.S., Tselykh A.N., Tselykh A.A. Application of cluster analysis for the assessment of the share of fraud victims among bank card holders // Proceedings of the 8th International Conference on Security of Information and Networks (September 08–10, 2015, Sochi, Russia). – 2015. – P. 103–106.
3. Alkhasov S.S., Tselykh A.N., Tselykh A.A. An Integrated ANN-GA Approach to Data Classification // Proceedings of the 2016 Conference on Information Technologies in Science, Management, Social Sphere and Medicine (ITSMSSM 2016). – 2016. – P. 172–176.
4. Алхасов С.С., Целых А.Н., Целых А.А. Классификация на основе модифицированной структуры искусственных нейронных сетей посредством генетических алгоритмов // Известия ЮФУ. Технические науки. – № 10 (183). – 2016. – С. 111–121.
5. Alkhasov S.S., Tselykh A.A. Combined Optimization and Modified Performance Metrics for Automated Model and Parameter Selection in Telecom Customer Churn Prediction // Proceedings of the IV Conference on Information Technologies in Science, Management, Social Sphere and Medicine (ITSMSSM 2017). – 2017. – P. 196–200.

Прочие труды

6. Алхасов С.С., Целых А.Н. Принципы построения прогностической системы для моделирования оттока клиентов // Материалы Всероссийской научной конференции «Системы и модели в информационную эпоху». Часть 1. – Таганрог: Изд-во ТТИ ЮФУ, 2014. – С. 4–6.
7. Алхасов С.С., Целых А.Н. Построение генетического алгоритма для использования в нейросетевом классификаторе // Современные информационные технологии и: тенденции и перспективы развития: Материалы XXIII научной конференции. – Ростов н/Д.: Изд-во ЮФУ, 2016. – С. 35–37.
8. Алхасов С.С., Целых А.Н. Основные элементы блока предварительной обработки результатов измерений в прикладных задачах анализа данных // Альманах современной науки и образования. – № 3 (105). – 2016. – С. 11–13.

9. Алхасов С.С., Целых А.Н., Целых А.А. Применение генетических алгоритмов со стохастической функцией приспособленности для оптимизации структуры нейронных сетей // Современные информационные технологии и: тенденции и перспективы развития: Материалы XXIII научной конференции. – Ростов н/Д.: Изд-во ЮФУ, 2016. – С. 44–46.
10. Алхасов С.С., Целых А.Н., Попкова Е.А. Решение задач классификации с использованием MATLAB и Microsoft Azure // Современные информационные технологии и: тенденции и перспективы развития: Материалы XXIII научной конференции. – Ростов н/Д.: Изд-во ЮФУ, 2016. – С. 38–43.

Свидетельство Роспатента

11. Алхасов С.С., Целых А.Н., Целых А.А. Свидетельство о государственной регистрации программы для ЭВМ №2016662656 от 17 ноября 2016 г. «Система классификаторов для прогнозирования оттока потребителей услуг телекоммуникационного предприятия».

Соискатель С.С. Алхасов