На правах рукописи

Орешков Вячеслав Игоревич

МЕТОДЫ И МОДЕЛИ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА ДАННЫХ В ЗАДАЧАХ УПРАВЛЕНИЯ В СОЦИАЛЬНЫХ И ЭКОНОМИЧЕСКИХ СИСТЕМАХ

Специальность 05.13.10 — Управление в социальных и экономических системах

АВТОРЕФЕРАТ

диссертации на соискание ученой степени

кандидата технических наук

Рязань — 2013

Работа выполнена в ФГБОУ ВПО «Рязанский государственный радиотехнический университет»
ФГБОУ ВПО «Рязанский государственный агротехнологический университет им. П.А. Костычева»

Научный руководитель: **ВАСИЛЬЕВ Евгений Петрович**
доктор технических наук, профессор,
ФГБОУ ВПО «Рязанский государственный агротехнологический университет им. П.А. Костычева», профессор кафедры экономической кибернетики.

Официальные оппоненты: **МАЛЫШ Владимир Николаевич**
доктор технических наук, профессор,
ФГБОУ ВПО «Липецкий государственный педагогический университет», заведующий кафедрой электроники, телекоммуникаций и компьютерных технологий.

МИТРОШИН Александр Александрович
кандидат технических наук, доцент,
ФГБОУ ВПО «Рязанский государственный радиотехнический университет», начальник управления телекоммуникаций и информационных ресурсов.

Ведущая организация: ФГБОУ ВПО «Владимирский государственный университет имени Александра Григорьевича и Николая Григорьевича Столетовых»

Защита состоится «27» июня 2013 г. в 12.00 на заседании диссертационного совета Д212.211.02 при ФГБОУ ВПО «Рязанский государственный радиотехнический университет» по адресу: 390005 г. Рязань, ул. Гагарина, д. 59/1.

С диссертацией можно ознакомиться в библиотеке ФГБОУ ВПО «Рязанский государственный радиотехнический университет».

Автореферат разослан «20» мая 2013 г.

Ученый секретарь диссертационного совета, канд. техн. наук  Перепелкин Д.А.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность работы. Ключевым фактором обеспечения качественного управления в социальных и экономических системах является организация непрерывного поиска новых, нетривиальных, практически полезных и доступных для интерпретации знаний, необходимых для эффективной поддержки принятия управленческих решений (УР). Важнейшим инструментом поиска знаний является глубокий и всесторонний анализ данных, описывающих процессы и явления в социальных и экономических системах, с использованием современных информационных технологий.

Высокая динамика и сложность современной экономической и социальной сфер предъявляет особые требования к организации таких исследований. Смещение центров принятия УР от высших эшелонов управления на уровень специалистов, непосредственно интегрированных в социальные, экономические и бизнес процессы, требует разработки методов и моделей анализа данных, которые могут применяться на практике широким кругом специалистов, не имеющими специального образования. Результаты анализа должны быть обобщаемы и тиражируемы для возможности применения построенных моделей для решения аналогичных задач на новых данных.

Наиболее перспективным направлением информационных технологий, используемым для организации поддержки принятия решений в социальных и экономических системах, в настоящее время является интеллектуальный анализ данных, также известный как Data Mining (DM) – раскопка, разработка данных. Это междисциплинарное направление, включающее элементы искусственного интеллекта (ИИ), математической статистики и машинного обучения (МО), применяемых для решения задач классификации, кластеризации и ассоциативного анализа.

Вместе с тем DM не дает шаблонов готовых решений и не предписывает строгих алгоритмов для той или иной задачи анализа. Он представляет собой методологию организации аналитической обработки данных, приемы и методы которой позволит извлечь из них максимум полезных знаний. Ядром аналитических технологий DM являются методы МО, позволяющие в автоматическом режиме восстанавливать структуры, зависимости и закономерности в данных, интерпретация и осмысление которых экспертом или аналитиком, позволяет делать заключении и выводы об особенностях состояния и развития явлений и процессов, вырабатывать рекомендации по более эффективному управлению ими.

Процесс внедрения DM-технологий в практическую деятельность предприятий и организаций для решения конкретных задач повышения эффективности управления в большинстве случаев достаточно затратный и трудоемкий. Основными проблемами являются отсутствие формальной постановки задачи и стратегии поиска знаний, эвристический характер большинства интеллектуальных моделей, высокая размерность и низкое качество данных. Поэтому разработка новых подходов и методов по реализации DM-проектов при решении конкретных задач повышения эффективности управления в социальных и экономических системах, является актуальной научно-технической задачей.

Степень разработанности проблемы. Развитие методов МО, как направлении ИИ связано с работами зарубежных ученых Б. Уидроу, М. Мински, П. Дж. Вербоса, Дж. Хопфилда, Д. Румельхарта, С. Пайперта, и отечественных: А.Б. Новикова, А. И. Галушкина, А.Н. Горбаня, С.И. Барцева, В.А. Охонина, В. Н. Вапника, А.Я. Червонескиса, Ю.И. Журавлева, К.В. Рудакова и др. В 70-80 г. XX в. в рамках МО были предложены деревья решений (Дж. Р. Куинлен, Л. Брейман), ассоциативные правила (Р. Агравал, Р. Шрикант), самоорганизующиеся карты признаков (Т. Кохонен) и др. Формирование DM как научного

направления связано с работами Г. Пятецкого-Шапиро, У. Файада, П. Смита и др. Значительный вклад в области моделирования социальных и экономических систем с целью анализа их функционирования и синтеза управленческих решений внесли В.Н. Бурков, Д.А. Новиков и др.

Предметом исследования в работе являются методы и алгоритмы DM, методология и проблемы их применения в задачах моделирования объектов и процессов в экономической, социальной и бизнес среде.

Объектом исследования избраны: аналитические технологии Data Mining, алгоритмы и методы МО: нейронные сети, деревья решений, карты Кохонена, ассоциативные правила, методы их применения для реализации практических задач анализа данных в социальных и экономических системах.

Цель работы. Разработка методов и моделей анализа данных в социальных и экономических системах с использованием интеллектуальных аналитических технологий Data Mining для повышения эффективности синтеза управленческих решений на основе знаний, обнаруженных в массивах данных.

Для реализации поставленной цели в диссертационной работе были поставлены и решены следующие задачи:

1) провести обзор и сравнительный анализ инструментальных средств DM и существующих подходов к организации процесса интеллектуальной аналитической обработки данных, разработать систему критериев и классификации аналитических инструментов;

2) определить основные факторы, влияющие на успешное внедрение аналитических DM-проектов на уровне специалистов, непосредственно интегрированных в процессы управления в социальных и экономических системах, разработана модель для оценки сложности аналитических DM-проектов;

3) разработать концепцию сценарного подхода к организации интеллектуальной среды аналитического DM-приложения на основе межотраслевого стандарта организации интеллектуального анализа данных CRISP-DM;

4) разработать сценарии построения базовых интеллектуальных моделей на основе нейронных сетей, деревьев решений, карт Кохонена, и интерфейс пользователя для их реализации;

5) разработать комплексную интеллектуальную модель урожайности зерновых по данным агрохимического обследования почв на основе нейронной сети, дерева решений, карт Кохонена и ассоциативной модели, агрегируемых в ансамбль на основе алгоритма стекинга;

6) разработать комплексную модель для анализа клиентской базы кредитной организации на основе ансамбля моделей, основанных на машинном обучении.

Соответствие паспорту специальности. Диссертационная работа выполнена в рамках п. 1.10 «Разработка методов и алгоритмов интеллектуальной поддержки принятия управленческих решений в экономических и социальных системах» и п. 1.12. «Разработка новых информационных технологий в решении задач управления и принятия решений в социальных и экономических системах», паспорта специальности 05.13.10 – «Управление в социальных и экономических системах». Теоретическую и методологическую основу исследования составили современная теория прикладной статистики, машинного обучения, искусственного интеллекта, теории информации, агротехнологий.

Информационно-эмпирическую базу исследований составили ведомости агрохимического обследования почв ОАО СПК «Рассвет» Тульской области и набор анкетных данных клиентов компании, специализирующейся в области потребительского кредитова-

ния. Обработка данных производилась на основе свободно распространяемой аналитической платформы Deductor Academic российской компании «ООО Аналитические технологии» (www.basegroup.ru).

Положения, выносимые на защиту и их научная новизна

1. Система классификации программных средств Data Mining с целью выбора программного обеспечения для реализации и внедрения проектов интеллектуального анализа данных. Существенными отличиями являются:

- максимально широкой охват инструментальных средств DM различных разработчиков и уровней сложности;

- разработка критериев и рекомендаций для выбора DM-средств с точки зрения внедрения на уровне специалистов, непосредственно интегрированных в процессы в социальных и экономических системах.

2. Двухуровневый сценарный подход к организации и управлению аналитическими проектами DM в области моделирования социальных и экономических систем в соответствии со стандартом CRISP-DM. Существенными отличиями от существующих подходов являются:

- иерархически структурированная последовательность операций аналитической обработки данных, представляемая в виде дерева с возможностью управления процессом моделирования посредством модификации его узлов и ветвей;

- сценарии построения интеллектуальных моделей, основанных на машинном обучении, с использованием декомпозиции процесса моделирования на этапы, реализуемые с помощью эвристических процедур;

- интеллектуальный интерфейс пользователя для реализации разработанных сценариев.

3. Комплексная модель урожайности зерновых по данным агрохимического обследования почв с помощью ансамбля интеллектуальных моделей, основанных на машинном обучении, агрегируемых с использованием стекинга. Основными отличиями являются:

- комплексное использование нескольких типов интеллектуальных моделей (нейронной сети, дерева решений, карты Кохонена и ассоциативной классификации) позволяет сопоставлять и сравнивать результаты, полученные с помощью различных моделей с целью оценки их согласованности и достоверности;

- концепция интеллектуального моделирования урожайности, позволяющая перейти от использования ретроспективных данных, к пространственным, что, в частности, более удобно для организации точного земледелия;

- усовершенствованный алгоритм построения дерева решений с автоматическим выбором наиболее значимого атрибута разбиения в условиях неопределенности критерия Gain-Ratio, на основе остаточной взаимной энтропии;

- усовершенствованная модель ассоциативной классификации на основе алгоритма поиска ассоциативных правил Apriori с использованием нового показателя - актуальности правил.

4. Комплексная интеллектуальная модель для анализа клиентской базы кредитной организации с целью совершенствования маркетинговой стратегии на основе исследования зависимости свойств клиента и его отклика на коммерческие предложения. Основными отличиями являются:

- комплексное применение нескольких моделей с целью повышения достоверности результатов и объясняющей способности бинарной классификации;

- методика сокращения размерности пространства входных признаков в условиях наличия большого количества числовых и категориальных факторов в исходных данных на основе применения дивергенции Кульбака-Лейблера.

Практическая значимость работы заключается в том, что сформулированные выводы и предложения, разработанные подходы и модели могут быть использованы широким кругом специалистов, занимающихся разработкой и внедрением DM-проектов на основе аналитических платформ и приложений. Модель оценки сложности аналитических проектов позволяет повысить эффективность планирования, разработки, реализации и внедрения проектов Data Mining. Модель урожайности на основе данных агрохимического обследования почв может быть использована предприятиями АПК, специализирующимися в области растениеводства, для повышения эффективности управления производством на основе оценивания урожайности с целью планирования севооборотов, оптимизации агротехнологических мероприятий и определения их экономического эффекта. Модель отклика клиентов на рекламную рассылку по анкетным данным может использоваться компаниями в области потребительского кредитования, для повышения эффективности маркетинговой стратегии и продвижения новых видов продуктов и услуг.

Апробация результатов работы. Основные результаты исследования докладывались и обсуждались на:

- Международной научно-практической конференции «Дни науки» (Прага, 2011);
- VII Международной научной конференции «Гуманитарные науки и современность» (Москва, 26 сентября 2012 г.);
- Всероссийской научно-практической конференции «Актуальные проблемы и их инновационные решения в АПК» (Рязань, 2011);
- Всероссийской-научно-практической конференции «Интеграция науки с сельскохозяйственным производством» (Рязань, 2011);
- семинарах и научных сессиях учетно-экономического факультета Рязанского государственного агротехнологического университета;
- семинарах и научных сессиях Рязанского государственного радиотехнического университета;
- результаты диссертационного исследования использовались в НИР "Разработка системы поддержки принятия решений в структурах АПК на основе современных платформ бизнес-аналитики", поддержанной субсидией Министерства сельского хозяйства и продовольствия Рязанской области на проведение работ по разработке приоритетных направлений научно-технического прогресса в агропромышленном комплексе.

Внедрение результатов исследования. Предложенные методы и модели аналитической обработки данных прошли успешную верификацию на реальных данных. Отдельные результаты диссертационного исследования нашли применение в практической деятельности компании ООО «НАНОАГРОТЕХ», ООО «Аналитические технологии». Результаты исследований применяются при чтении курсов лекций «Информационные технологии в экономике», в Рязанском государственном агротехнологическом университете, «Интеллектуальные подсистемы САПР» в Рязанском государственном радиотехническом университете, «Статистика» по специальности «Государственное и муниципальное управление» и «Управление персоналом» в Рязанском государственном университете им. С.А. Есенина.

Публикации. По теме диссертации опубликовано 18 работ, в том числе: 6 статей в изданиях, рекомендованных ВАК РФ, 1 монография (2 издания: 2009 и 2011 г.), 1 учебное пособие, 10 работ в изданиях, зарегистрированных в Госкомнадзоре РФ и сборниках трудов научных и научно-практических конференций.

Структура и объем работы. Диссертация состоит из введения, 4-х глав, заключения, списка литературы и 3 приложений, которые содержит документы о внедрении и практи-

ческом использовании полученных результатов, таблицы исходных данных и интерфейсы. Основной текст работы содержит 209 страниц, 76 рисунков, 31 таблицу. Список литературы включает 127 наименований.

СОДЕРЖАНИЕ РАБОТЫ

Во введении обосновывается актуальность выбранной темы, определяются цели и задачи, рассматриваемые в диссертационной работе, перечислены полученные новые научные результаты, сформулированы основные положения, выносимые на защиту, представлены ее практическая ценность и апробация.

Первая глава посвящена обзору инструментальных средств, технологий и методов реализации аналитических проектов Data Mining, а также разработке критериев и системы классификации программных продуктов Data Mining с целью выработки рекомендаций по их выбору для решения задач поиска знаний в социальных и экономических системах.

К концу первого десятилетия XXI в. рынок аналитического ПО, использующего технологии DM, достиг объема 7,8 млрд. долл. США (с ежегодным ростом 12,1%). Крупнейшими поставщиками решений на рынке ПО для DM стали: SAS Institute (SAS Enterprise Miner – 33,2%), IBM (IBM SPSS Modeler – 14,3%, до 2009 г. SPSS Clementine), Microsoft (SQL Server Analysis Services, 1.7%), Teradata (TeraMiner, 1.5%), and TIBCO (TIBCO Spotfire, 1.4%).

С середины 1990 г. популярными становятся библиотеки с открытым исходным кодом (WEKA, XELOPES). Большую группу DM-инструментов образуют так называемые прототипы – системы компьютерной математики изначально не ориентированные на DM, но содержащие операторы и функции, поддерживающие реализацию алгоритмов и методов ИАД (тулбоксы MATLAB, библиотеки языка R и т.д.). Параллельно с ростом числа доступных DM-инструментов росла их сложность для большинства потенциальных пользователей, а также обострялась проблема выбора наиболее подходящего продукта. Поэтому актуальной задачей является разработка критериев и системы классификации DM-продуктов для их обоснованного выбора для реализации аналитических проектов в социальных и экономических системах. На основе анализа целей, задач и практических реализаций различных DM-проектов автором разработана система критериев для классификации DM-приложений (таблица 1).

Таблица 1 - Классы DM-приложений

Обозначение	Описание
DMST – Data Mining Suite Tools (аналитические платформы)	Содержат множество методов и алгоритмов анализа и моделирования, поддерживают работу с многомерными структурированными и неструктурированными данными, не являются проблемно-ориентированными, включают весь спектр функций DM, необходимых для создания завершенных аналитических проектов: интегрирование с бизнес-приложениями, импорт/экспорт данных и компонентов, формирование аналитической отчетности.
DMBT - Data Mining Business Tools (пакеты бизнес-аналитики)	Не создавались изначально для решения задач DM, но включают его отдельные элементы: статистические методы, средства формирования аналитической отчетности и др. Имеют высокую интегрируемость в бизнес-структуры, и возможности работы с разнообразными источниками данных.
DMMP - Data Mining Mat Package (СКМ с элементами DM)	Системы компьютерной математики с элементами DM, содержат алгоритмы и средства визуализации, позволяющие реализовывать функциональность DM, работать с изображениями, видео и звуковыми файлами. Интерактивность реализуется с помощью встроенного языка программирования.
IDMT – Integration Data Mining Tool (интегрируемые DM-приложения)	Наборы алгоритмов DM, образующих отдельные программные средства, либо пакеты расширения. Являются средствами разработки, не имеют графического интерфейса, возможности по очистке и предобработки данных ограничены.
DMEP – Data Mining Extend Package (пакеты расширения DM)	Модули подключения к Excel, Matlab и другим приложениям, реализующие определенную (как, правило, узкую) функциональность DM. Не имеют собственного интерфейса пользователя, а также средств экспорта/импорта данных.

Продолжение таблицы 1

DMLT - Data Mining Library Tools (библиотеки функций DM)	Наборы функций DM, которые могут быть внедрены в другие приложения с помощью API. Графический интерфейс отсутствует, поэтому используются они в основном разработчиками.
SDMT – Specialties Data Mining Tools (специализированные средства DM)	Средства, ориентированные на использование какого-либо одного семейства алгоритмов или методов DM - нейронных сетей, деревьев решений, ассоциативных правил и т.д.
RDMT – Research Data Mining Tools (исследовательские средства DM)	Экспериментальное ПО, включающее новые, экспериментальные алгоритмы и исследования в области DM. Являются средством разработки, не имеют графического интерфейса, развитых средств очистки, экспорта и импорта данных.
DMFT – Data Mining Field Tools (проблемно - ориентированные DM-приложения)	Средства, ориентированные на определенную прикладную область, например, анализ текста (TextMining), анализ мультимедиа-данных, анализ геоинформационных систем (Spatial Data Mining).

Анализ классов аналитического ПО в совокупности с требованиями, предъявляемыми к DM-продуктам, ориентированным на создание завершенных аналитических проектов масштаба предприятия, позволяет произвести сравнение перечисленных классов аналитического ПО с точки зрения перспективности их использования для внедрения на уровне специалистов, непосредственно интегрированных в социальные, экономические и бизнес-процессы. Результаты сравнения представлены в таблице 2.

Таблица 2 – Сравнение классов аналитического ПО Data Mining

Класс приложения	Экспорт/ импорт данных	Наличие GUI	Очистка и трансформация	Разнообразие алгоритмов и методов	Визуализация	Технология клиент-сервер	Аналитическая отчетность	Средство разработки	Практический анализ
DMST	+	+	+	+	+	+	+		+
DMBT	+	+	+			+	+		+
DMMP	+			+				+	
IDMT		+	+	+				+	
DMEP									+
DMLT								+	
SDMT	+	+	+		+				+
RDMT								+	
DMFT	+	+	+		+				+

На основе результатов сравнения можно сделать вывод, что наиболее подходящим классом ПО для реализации и внедрения аналитических проектов DM масштаба предприятия аналитические платформы (DMST), содержащие весь комплекс средств, необходимых для организации процесса поиска и тиражирования знаний. К данному классу относятся зарубежные коммерческие продукты Estard Data Miner, Deductor Enterprise Miner, SAS Enterprise, Miner, ISOFT Alice, DataEngine, DataDetective, GhostMiner, Knowledge Studio, KXEN, Partek Discovery Suite, а также отечественные разработки PolyAnalyst (Мегапьютер Интиллидженс) и Deductor (ООО «Аналитические технологии»).

Во второй главе производится разработка аналитической среды, обеспечивающей максимально эффективную работу широкого круга специалистов, непосредственно интегрированных в социальные, экономические и бизнес-процессы, с методами и моделями Data Mining. Базовой идеологией формирования такой среды является исключение необходимости понимания пользователем математических аспектов построения моделей и технических аспектов

средств управления данными, что даст ему возможность сфокусироваться на решении задач интерпретации результатов и синтезе управленческих решений.

В основе построения такой среды лежит декомпозиция сложных операций аналитической обработки данных на последовательность простых действий, каждое из которых может быть выполнено на основе эвристических правил и рекомендаций. Структурированную, формализованную, и описанную последовательность таких действий будем называть сценарием. Тогда DM-проект можно рассматривать как набор сценариев, применяемых к одному или нескольким источникам данных и реализующих определенную процедуру их обработки. Разработанный и проверенный сценарий сохраняется пользователем в специальном файле проекта, откуда может быть впоследствии вызван для применения новых данных.

Предлагаемый сценарный подход можно рассматривать как альтернативу поточному (data stream, knowledge flow) подходу, который является доминирующим способом формирования аналитической среды в зарубежных DM-платформах. При использовании поточного подхода в рабочем поле приложения размещаются пиктограммы операторов, реализующих определенные функции Data Mining и управления данными, выбираемые пользователем из библиотеки. Затем производится настройка параметров каждого оператора, и они соединяются стрелками, указывающими путь прохождения данных. При этом иерархия, структурированность и логическая последовательность операторов не контролируется системой, что затрудняет не только разработку новых процедур обработки данных, но и понимание существующих. Такой подход требует от пользователя достаточно высокого уровня знаний в области анализа данных и не является эффективным при реализации DM-проектов на уровне «массового» пользователя.

В основе сценарного подхода лежит идея движения не от действия (оператора), а от задачи: пользователь формулирует задачу, а система сама «подсказывает» варианты и последовательность действий для ее решения. Движение «от задачи» является более эффективным, поскольку набор задач, реализуемых в процессе разработки DM проекта является стандартным и включает одни и те же шаги для проектов в различных проблемных областях и регламентируется Межотраслевым стандартом обработки данных для Data Mining (Cross Industry Standard Process for Data Mining – CRISP-DM). Обобщенная структурная схема DM, разработанная на основе стандарта CRISP-DM, проекта представлена на рис. 1.

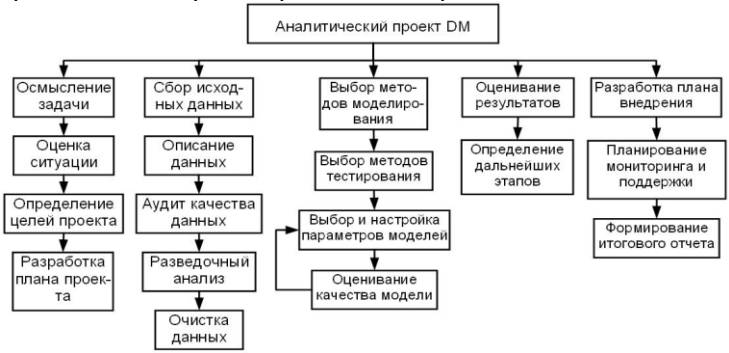


Рис. 1 – Обобщенная структурная схема аналитического DM-проекта на основе стандарта CRISP DM.

Аналогичным образом можно выполнить и декомпозицию процедуры построения интеллектуальных моделей. Автором выполнена разработка сценариев для базовых анали-

тических DM-моделей, входящих в состав большинства аналитических платформ – нейронных сетей, деревьев решений и карт Кохонена (рис. 2).

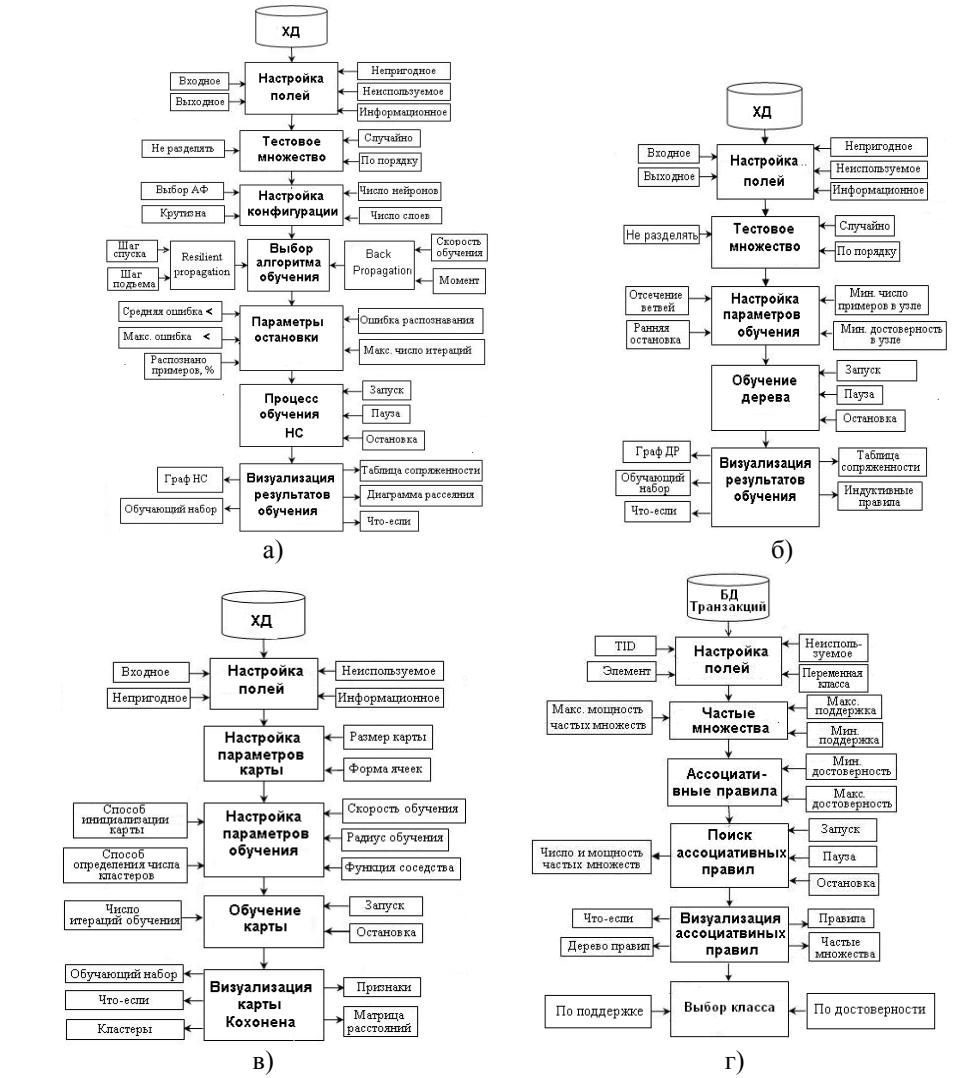


Рис. 2 - Сценарии построения аналитических моделей: а) нейронной сети, б) дерева решений, г) карт Кохонена, д) ассоциативной модели.

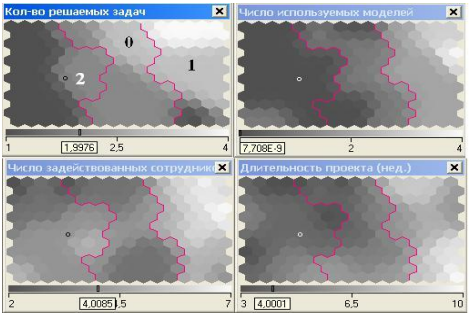
Важной проблемой, решаемой в процессе реализации DM-проектов, является планирование ресурсов – времени, требуемого на реализацию проекта, числа задействованных сотрудников, число задач, которое требуется решить, количества моделей, которое требуется построить. Поэтому практический интерес представляет разработка методики оцени-

вания сложности аналитического проекта. Данная задача является плохо формализованной, поскольку строго обоснованные критерии сложности проекта отсутствуют. Кроме этого в процессе реализации проекта возникают условия, трудно поддающиеся учету. Например, компания заказчик задержала исходные данные, что привело к увеличению сроков проекта, или в процессе работы над проектом выяснилось, что число задач, которые требуется решить, больше запланированного.

Поэтому для оценки сложности DM проектов предложен подход на основе использования интеллектуальных моделей, основанных на машинном обучении. Была собрана информация о 52 проектах, реализованных на основе аналитической платформы Deductor. На ее основе был сформирован обучающий набор данных, содержащий следующие признаки: отрасль, в которой выполнялся проект; количество задач, решаемых в проекте, число используемых для этого аналитических моделей, число задействованных сотрудников и срок завершений проекта (недель).

Поскольку целевая переменная отсутствует, предварительно необходимо выполнить группировку похожих проектов и попытаться ассоциировать их с уровнями сложности. Для этого использовалась кластеризация на основе карт Кохонена при числе кластеров, равном 3. Построение карты производилось в соответствии со сценарием, представленным на рис. 2, в. Карты, построенные по каждому признаку, и параметры обучения, представлены на рис. 3, а.

Согласно алгоритму построения карты светлые ячейки соответствуют бо́льшим значениям признака, а темные – меньшим. Следовательно, кластер №1 содержит проекты с наибольшими значениями признаков, и его можно ассоциировать с уровнем сложности «Высокий», кластер №2 – «Низкий», а №0 – «Средний». Данные метки класса были присвоены проектам, попавшим в соответствующие кластеры, что позволило сформировать обучающее множества для построения классификационной модели на основе дерева решений. Правила, на естественном языке, извлеченные из ДР, представлены на рис. 3, б.



а)

Номер правила	Условие			Следствие	Поддержка	
	Показатель	Знак	Значение	Класс	Кол-во	%
1	Число используемых моделей	<	1,5	Низкая	21	40,38
	Кол-во решаемых задач	<	1,5			
6	Число используемых моделей	>=	1,5	Высокая	13	25,00
5	Число используемых моделей	<	1,5	Средняя	8	15,38
	Кол-во решаемых задач	>=	1,5			
	Число используемых моделей	>=	0,5			
3	Число используемых моделей	<	1,5	Низкая	4	7,69
	Кол-во решаемых задач	>=	1,5			
	Число используемых моделей	<	0,5			
	Длительность проекта (нед.)	<	5,5			
	Число задействованных сотруд	>=	3,5			
2	Число используемых моделей	<	1,5	Средняя	4	7,69
	Кол-во решаемых задач	>=	1,5			
	Число используемых моделей	<	0,5			
	Длительность проекта (нед.)	<	5,5			
	Число задействованных сотруд	<	3,5			
4	Число используемых моделей	<	1,5	Средняя	2	3,85
	Кол-во решаемых задач	>=	1,5			
	Число используемых моделей	<	0,5			
	Длительность проекта (нед.)	>=	5,5			

б)

Рис. 3. Карта Кохонена (а) и дерево решений (б), используемые для оценки сложности DM-проектов.

Наиболее значимым является правило №1 :«**Если** число используемых моделей меньше двух и количество решаемых задач меньше двух, **то** класс сложности проекта – низкий». Правило № 6 позволяет классифицировать все проекты высокой сложности и утверждает, что такие проекты должны использовать 2 или более интеллектуальные модели. И, наконец, правило №8, классифицирующее большую часть проектов с меткой «Средняя» утверждает, что проект

имеет среднюю сложность, если число используемых моделей равно 1, но число задач 2 и более. Данные 3 правила позволяют классифицировать более 75% проектов.

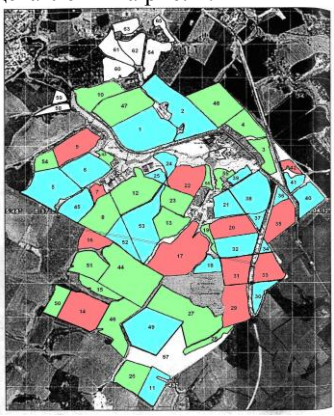
В третьей главе произведена разработка и построение комплексной модели урожайности на основе данных агрохимического обследования почв. Топографическая основа и фрагмент ведомости агрохимического обследования представлены на рис. 4.

№ поля	Площадь, га	Кислотность, pH	Азот, мг/100г	Калий, мг/100г	Фосфор, мг/100г	Юг, %	Ср. угол	Урожайность	Урожайность (класс)
1	183,01	4,8	3	19,3	18,52	91,36	1,28	2,2	Низкая
2	185,99	5,5	5	10,83	19,15	67,9	0,83	4,8	Средняя
3	47,52	5,6	6	12,76	16	44,24	0,92	6,3	Средняя
4	63,04	5,6	6	12,76	16	92,8	0,69	6,4	Средняя
5	141,08	5,2	4	24,42	19,55	26,26	1,05	3,5	Низкая
6	100,06	5,2	4	14,89	10,8	20,05	1,51	3,4	Низкая
7	18,15	5	3	7,4	6	1,33	1,78	2,2	Низкая
8	99,64	5,4	5	15,97	25,2	0,12	1,38	4,7	Средняя
9	91,36	5,1	4	9,31	6,94	12,38	1,36	3,5	Низкая

а)

Показатель	Уравнение регрессии	Статистика R^2	Коэффициент корреляции, R	Ожидаемый прирост при увеличении на единицу (ц/100га)
Пашня с уклоном к югу, %	$y=0{,}01x+4{,}165$	0,01	0,23	1,0
Площадь поля, га	$y=0{,}005x+4{,}067$	0,01	0,30	0,5
Средний угол уклона, °	$y=0{,}6x+3{,}78$	0,01	0,30	6,0
Содержание калия, мг/100г	$y=0{,}022x+1{,}813$	0,13	0,36	2,2
Содержание фосфора, мг/100г	$y=0{,}316x-0{,}16$	0,34	0,58	31,6
Кислотность, pH	$y=2{,}035x-6{,}245$	0,38	0,61	203,5
Содержание азота, мг/100г	$y=0{,}966x-0{,}418$	0,71	0,83	96,6

б)



в)

Рис. 4. а) фрагмент ведомости агрохимического обследования полей, б) результаты разведочного анализа, в) топографическая основа.

Целями моделирования являются: 1) исследование зависимости урожайности от агрохимических параметров с целью оптимизации планирования агрохимических и агротехнических мероприятия; 2) предсказание урожайности для новых полей с целью планирования севооборотов.

Исходные данные представляют собой ведомость агрохимического обследования почв СПК «Рассвет» Тульской обл. по яровому ячменю. Всего обследовано 64 поля, из которых фактическая урожайность известна для 56, остальные 8 полей использовалась для верификации моделей. Ведомость содержит следующие характеристики: площадь поля (га), процент пашни с уклоном к югу, средний угол уклона, содержание макроэлементов (мг/100 г почвы): азота, калия и фосфора, а также кислотность почв pH, фактическая урожайность (ц/га). Диапазон изменения наблюдаемой урожайности был разделен на три равных поддиапазона [0,4[- «Низкая», [4,8[- «Средняя» и [8,12] – «Высокая». Таким образом, урожайность можно представить с помощью лингвистической переменной:

$$\{y, Y, T(y), G, M\},$$

где: y= «Урожайность» - наименование переменной; Y=[0,12] – множество значений переменной Y (вещественных чисел в диапазоне от 0 до 12); $T(y)$ – {«высокая», «средняя», «низкая»} – терм-множество; $G(y)$ – {«очень», «не очень»} – синтаксическое правило; M – семантическая процедура.

Разведочный анализ данных. Важным этапом проекта DM является разведочный анализ (РА). Его задача – определить характер и структуру данных, общий вид и логику зависимостей между переменными, оценить их значимость для решения задачи. Обычно РА производится на основе простых статистических методов, таких как корреляционный и регрессионный анализ, анализ трендов и т.д. Если размерность данных не высока, полезным оказывается визуальный анализ графиков и таблиц.

Основной целью РА является оценка целесообразности использования для моделирования имеющихся показателей. Корреляционно-регрессионный анализ (рис. 4, б), показал, что процент пашни с уклоном к югу, средний угол уклона и площадь поля практически не обеспечивают прироста урожайности. Кроме этого, изменение этих параметров на практике не реализуемо. Наибольший прирост урожайности обеспечивают кислотность и содержание азота, поэтому поиск их связи с урожайностью представляет большой практический интерес. Хотя содержание фосфора и калия и не дают значительного прироста, тем не менее, их целесообразно включить в модель, поскольку они, наряду с содержанием азота, являются важными факторами обеспечения растений питательными веществами.

Модель множественной линейной регрессии урожайности по выбранным характеристикам имеет вид:

$$y = 0{,}34 - 0{,}84(Кислотность) + 1{,}1(Азот) + 0{,}16(Фосфор) + 0{,}08(калий)$$

Модель обеспечивает среднеквадратическую ошибку оценивания 1,3 ц/га, что составляет более 10% от наблюдаемого диапазона урожайности.

Для построения комплексной модели урожайности использовались нейросетевая модель (НСМ), дерево решений (ДР), карта Кохонена (КК) и ассоциативная модель (АМ). Построение моделей производилось по сценариям, разработанным в гл. 2. Применение данного комплекса моделей обеспечит оценку урожайности методом численного предсказания, классификации, кластеризации и ассоциации. Это позволит повысить интерпретируемость и достоверность результатов итоговой метамодели.

Построение НСМ. Построение НСМ производилось на основе сценария, разработанного в гл. 2 (рис. 2,а). В процессе построения модели решались следующие задачи.

1. Определении базовой архитектуры и конфигурации сети. Была выбрана архитектура плоскослоистой НС с последовательными связями и сигмоидальной активационной функцией (АФ) (персептрон Румельхарта). Использовалась конфигурация с одним скрытым слоем число нейронов в котором определялось по эвристическому правилу – число связей в сети в 2-3 раза меньше числа обучающих примеров, применение которого снижает эффект переобучения (выбор проверен экспериментально). Произведен экспериментальный выбор АФ и параметра крутизны. Результаты представлены на рис. 5.

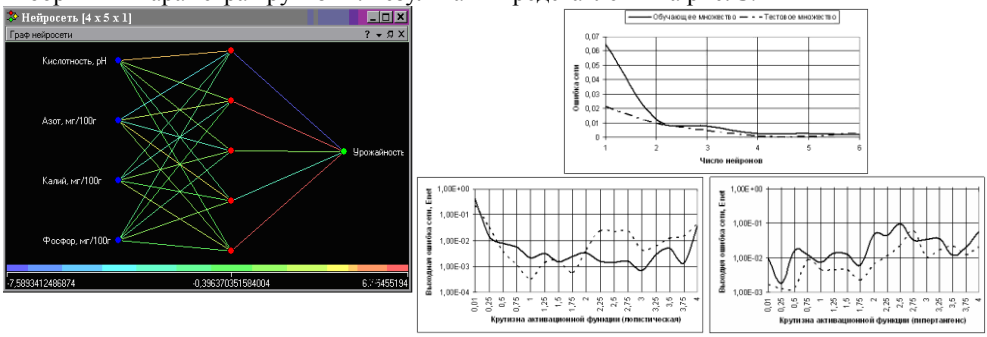
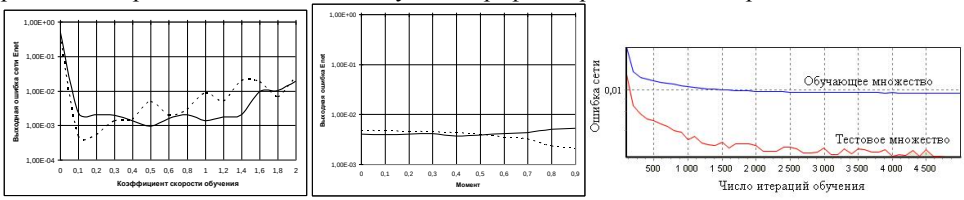


Рис. 5 - Граф НСМ и графики экспериментов по выбору параметра крутизны АФ и числа нейронов в скрытом слое.

Таким образом, была выбрана конфигурация НСМ с 5 нейронам в скрытым слое, логистической активационной функцией и параметров крутизны α=3.

2. Выбор параметров алгоритма обучения. Для обучения НСМ был выбран алгоритм обратного распространения ошибки, как наиболее устойчивый к низкому качеству исходных данных. Параметры алгоритма (коэффициент скорости η=0,5 обучения и момент μ=0,1) подбирались экспериментально. Соответствующие графики представлены на рис. 6.

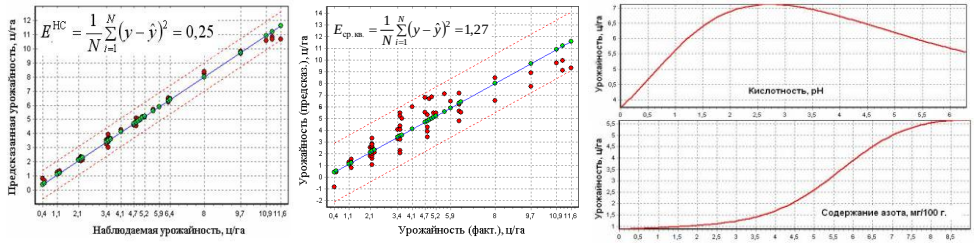


а) б) в)

Рис. 6 – Выбор параметров обучения НСМ: а) коэффициента скорости обучения, б) момент, в) зависимости ошибки сети от числа итераций.

График ошибки обучения на обучающем и тестовом множестве показал, что ошибка сети перестает уменьшаться по достижении примерно 2000 итерации, поэтому дальнейшее обучение НСМ не имеет смысла.

3. Оценка результатов обучения НСМ. Для визуальной оценки результатов обучения НСМ воспользуемся диаграммой рассеяния (рис. 7, а) и сравним ее с диаграммой рассеяния для множественной линейной регрессии (рис. 7, б). Рассеяние оценок относительно линии $y = \hat{y}$ для НСМ намного меньше, чем для регрессии. Среднеквадратическая ошибка НСМ $E^{HC} = 0{,}25$ ц/га почти в пять раз ниже, чем для регрессии $E_{ср.кв} = 1{,}27$.



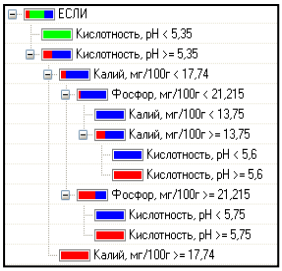
а) б) в)

Рис. 7 – Диаграммы рассеяния: а) НСМ, б) множественной регрессии; в) диаграммы «что-если» для кислотности и содержания азота.

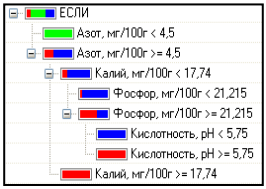
4. Разработка методики применения НСМ. Для практического использования НСМ с целью планирования агрохимических мероприятий использовался анализ «что-если» На рис. 8, в) представлены диаграммы «что-если» для кислотности и азота. Пусть на рассмотрение поступает новое поле с характеристиками: pH=4,9, Азот=3 мг/100 г., Калий=18,2 мг/100 г., Фосфора=16,42 мг/100 г. и предсказанной урожайностью 2,36 ц/га. Анализ графиков показывает, что возможности увеличения урожайности по параметрам pH и Калий исчерпаны. Увеличение содержания азота с 3 до 8 мг/100 г. (при фиксации значений остальных параметров) потенциально позволит увеличить урожайность до 5,5 ц/га. Поэтому в данном случае предпочтительным является снесение азотных удобрений.

Построение модели урожайности на основе ДР. Построение ДР производилось на основе сценария, разработанного во 2 гл (рис. 2, б) с помощью алгоритма ID3, который на каждой итерации генерирует правило, позволяющие сформировать узлы с минимальной

энтропией (т.е. максимально однородные по классовому составу). Дерево, сформированное алгоритмом, и извлеченные из него правила, представлены на рис. 8.



а)



б)

№	Условие			Следстве	Поддержка		Достоверность	
	Показатель	Знак	Значение	Урожай	Кол-во	%	Кол-во	%
1	Азот, мг/100г	<	4,5	Низкая	30	53,57	30	100,00
2	Азот, мг/100г	>=	4,5	Средняя	17	30,36	16	94,12
	Калий, мг/100г	<	17,74					
	Фосфор, мг/100г	<	21,215					
3	Азот, мг/100г	>=	4,5	Средняя	2	3,57	2	100,00
	Калий, мг/100г	<	17,74					
	Фосфор, мг/100г	>=	21,215					
	Кислотность, pH	<	5,75					
4	Азот, мг/100г	>=	4,5	Высокая	3	5,36	3	100,00
	Калий, мг/100г	<	17,74					
	Фосфор, мг/100г	>=	21,215					
	Кислотность, pH	>=	5,75					
5	Азот, мг/100г	>=	4,5	Высокая	4	7,14	4	100,00
	Калий, мг/100г	>=	17,74					

в)

Рис. 8 – Результаты построения ДР: а) ДР построенное с помощью стандартного алгоритма ID3, б) ДР построенное на основе модифицированного алгоритма, в) правила, извлеченные из оптимизированного ДР.

Недостатком ДР на рис. 8,а является отсутствие в правилах признака «Азот», в то время, как РА и НСМ показали его как наиболее важный для урожайности признак. Исследование проблемы показало, что в классическом алгоритме ID3 отсутствует автоматическая обработка ситуации, когда нормированный прирост информации $GR = IG_i / IV$ (где IG_i (information gain) – увеличение информации при разбиении по i-му атрибуту, IV (intrinsic value) – полный прирост информации при разбиении) на основе максимального значения которого выбирается атрибут ветвления в узле, оказывается одинаковым для двух или более атрибутов . Это вызывает неопределенность выбора. В различных приложениях данная проблема решается путем выбора первого встретившегося атрибута, или выбор предоставляется пользователю. Ни тот, ни другой вариант не является оптимальным – ДР может стать слишком большим и сложным, из него могут «выпасть» правила, представляющие значительный интерес с точки зрения логики анализа.

Для решения проблемы предлагается модифицировать алгоритм путем ввода в него дополнительного критерий выбора атрибута ветвления, используемый в описанном выше случае, - взаимную остаточную энтропия между атрибутом, который является кандидатом для формирования правила в узле:

$$H_a(x,c) = -\sum_k \sum_j p(x_j c_k) \log_2 p(x_j c_k),$$

где $p(x_j c_k)$ - совместная вероятность появления j-го значения атрибута x и k-го значения переменной класса для примеров не распределенный ни в один узел. Иными словами, алгоритм всегда будет отдавать предпочтение тем атрибутам, для которых изменчивость переменной класса ниже, что позволит ему формировать более однородные в смысле классового состава подмножества и завершать построение дерева за меньшее число шагов.

Применение модифицированного алгоритма позволило получить дерево, представленное на рис. 9, б. Оно компактнее исходного дерева (8 узлов вместо 12 и 5 правил вместо 7) и содержат все доступные показатели, что увеличивает объясняющую способность модели. Анализ правил оптимизированного дерева (рис. 8, в) показывает, что низкое (менее 4,5 мг/100 г) содержание азота в почве всегда связано с низкой урожайность (правило №1), а

высокая урожайность имеет место при содержании азота, большем 4,5 мг/100 г и низкой кислотностью pH>5,75. Это совпадает с результатами, полученными на основе НСМ, и соответствует теории растениеводства.

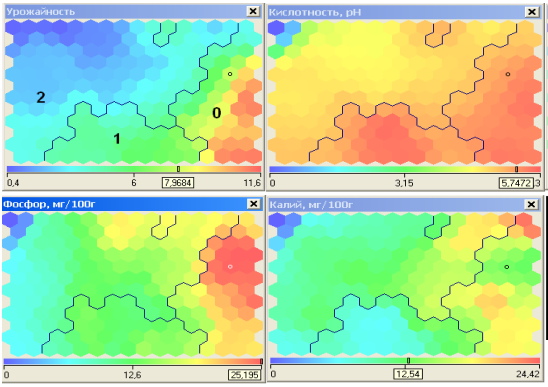
Построение кластерной модели урожайности на основе карты Кохонена. Построение и обучение карты Кохонена производилось в соответствии со сценарием, разработанным в гл. 2 (рис. 2.3). В процессе моделирования были решены следующие задачи:

1. Выбор входных признаков для кластеризации. Поскольку КК реализуют парадигму обучения без учителя, фактическая урожайность в качестве переменной класса не требуется и ее можно использовать как дополнительную входную перемнную, что позволит улучшить качество кластеризации.

2. Выбор параметров карты. Число ячеек карты выберем в 2 – 3 раза больше, числа обучающих примеров (16x18=192 ячейки), форма ячеек – шестиугольная.

3. Выбор параметров обучения карты. Количество итераций обучения – 1000, способ начальной инициализации – случайными значениями, параметр скорости обучения 0,5 – 0,05, радиус обучения 3,0 – 0,1, функция соседства гауссова. Количество кластеров выберем равное числу уровней урожайности, исходя из гипотезы, что поля с различными уровнями урожайности образуют устойчивые группы.

4. Содержательная интерпретация построенных карт (рис. 9)



	По кластеру			
Факт	0	1	2	Итого
Высокая	**8**		1	9
Средняя	3	**14**		17
Низкая	1		**29**	30
Итого	12	14	30	56

Рис. 9 – Карты Кохонена, построенные по результатам кластеризации.

Отфильтровав поля по кластерам, получим следующее распределение. Из 12 объектов, попавших в кластер № 0, 8 объектов имеют класс «Высокая», 3 – «Средняя» и 1 - низкая, что позволяет ассоциировать кластер с классом «Высокая». Из 14 объектов, попавших в кластер №1 все 14 относятся к классу «Средняя», а из 30 объектов в кластере № 2 – 29 - к классу «Низкая», что позволяет ассоциировать кластеры с соответствующими уровнями урожайности.

Моделирование урожайности на основе ассоциативного анализа. Ассоциативная модель (АМ) представляет собой систему правил «из А следует В» (записывается А→В), где А (условие) и В (следствие) – события, явления или объекты (или их подмножества), появляющиеся совместно. Чтобы использовать АМ для классификации, нужно перейти от обычных ассоциативных правил (АП) к классифицирующим ассоциативным правилам (КАП):

$$A_1, A_2, \ldots, A_m \to C_j \cdot \; j = 1..k \,, \qquad (1)$$

k - число классов. Поскольку АП работают с категориальными величинами, для формирования БД транзакций необходимо преобразовать исходные признаки к интервальным значениям и сформировать для них мнемонические метки. Это можно сделать с помощью таблиц обеспеченности почвы питательными веществами (табл. 3)

Таблица 3. Классы кислотности почв и обеспеченности питательными веществами

Группы	Азот мг/100 г.	P_2O_5 мг/100 г.	K_2O мг/100 г	Зерновые
1.	<5	< 2	< 2,0	Оч. низкая
2.	5-9	2–5	2–4	Низкая
3.	9-20	5–10	4–8	Средняя
4.	20-30	10–15	8–12	Повыш.
5.	30-40	15-20	12–18	Высокая
6.	40>	> 20	> 18,0	Оч. высокая

Класс	Кислотность почв	
	степень	*pH*
1	Оч.сильнокислая	< 4,0
2	Сильнокислая	4,1–4,5
3	Среднекислая	4,6–5,0
4	Слабокислая	5,1–6,0
6	Нейтральная	6,1–7,0
7	Щелочная	> 7,0

Пример подобной транзакции:

$$Среднекислая, Азот\_Высок, Калий\_Оч\_Низк, Фосфор\_оч\_низак \rightarrow Средняя$$

Затем с помощью стандартного алгоритма Apriori производится поиск КАП вида (1) и для каждого вычисляется поддержка и достоверность

$$Sup(C_j) = \frac{N(A_1A_2...A_n \cup C_j)}{N} \qquad Conf(C_j) = \frac{N(A_1A_2...A_n \cup C_j)}{N(A_1A_2A_n)}.$$

Решающим правилом для выбора класса будут $C = \max_j \{Sup(C_j)\}$ или. $C = \max_j \{Conf(C_j)\}$.

Для ограничения числа правил, порог поддержки обычно выбирают достаточно большим (0,5 и выше). Тогда правила с редкими классами, встречающимися менее чем в 50% примерах, не будут обнаружены алгоритмом. В соответствии со свойством антимонотонности, лежащим в основе поиска АП, модель будет классифицировать только примеры с классом «Низкая», т.к. поддержка любых ассоциаций с классами «Средняя» и «Высокая» не будет превышать 0,32 и 0,14 соответственно. В то же время, с точки зрения логики задачи, наибольший интерес представляют как раз поля с высокой урожайностью и факторы ее обуславливающие. Чтобы автоматизировать обнаружение редких классов, автором введена новая мера, - значимость АП. отношение частоты появления условия и следствия (т.е. поддержки ассоциации в целом $S(A \rightarrow B)$ к частоте появления только следствия $S(B)$), т.е. $R = S(A \rightarrow B)/S(B)$. Актуальность правила это безразмерная величина, которая изменяется в диапазоне от 0 до 1. При этом, чем выше актуальность, тем выше потенциальный интерес правила для аналитика, несмотря на его низкую поддержку. Актуальность позволяет обнаруживать интересные, с точки зрения логики решаемой задачи, ассоциации, даже если соответствующие правила не удовлетворяют условиям минимальной поддержки и достоверности и отбрасываются алгоритмом как малозначимые. Для этого правила нужно ранжировать по убыванию поддержки. По мере того, как поддержка монотонно убывает, значимость будет давать «всплеск» при появлении в правилах нового класса (рис. 10)

Рис. 10 - Графики актуальности и достоверности АП.

Актуальность АП позволяет не только обнаруживать редкие классы, но и выполнять балансировку модели по правилу $S_i^{\text{корр}} = (1-\alpha(1-R_i))\cdot S_i$. При α=0 модель лучше обнаруживает классы с высокой поддержкой. При α=1 на поддержку правил с частыми классами накладывается штраф. Эксперименты показали, что подбор параметра α позволяет снизать ошибку классификации модели на 12-15%.

В четвертой главе произведена разработка интеллектуальной модели клиентской базы данных (КБД) кредитной организации с целью уточнения целевой аудитории рекламных акций и снижения расходов на нее. КБД содержит информацию о 15244 клиентах по 42 признакам, из которых 13 целого типа и 39 строкового. Целевая переменная – переменная отклика на рекламную рассылку, принимающая значение 1 (положительный исход, 1812 записей), если отклик имел место, и 0 в противном случае (отрицательный исход, 13411 записей). Таким образом, вероятность отклика при рассылке всем клиентам не превышает 11%.

Целями моделирования являются: (1) предсказание реакции клиента на коммерческое предложение с целью принятия решения о целесообразности контакта; (2) выявление факторов, влияющих на восприимчивость клиентов к новым услугам, предлагаемым компанией, с целью уточнения целевой аудитории и снижения расходов на рассылку. Структурная схема модели представлена на рис. 11.

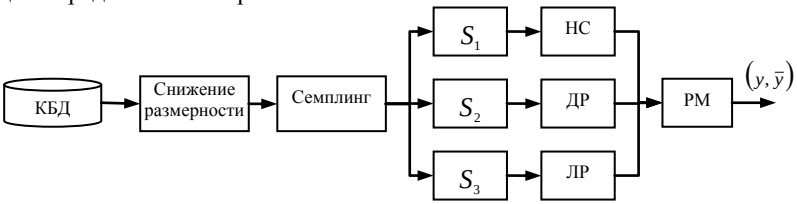


Рис. 11 – Структурная схема процедуры анализа клиентской базы.

На схеме обозначено:

- снижение размерности – применяется алгоритм снижения размерности исходных данных путем отбора значимых признаков;

- семплинг – процедура отбора записей в подмножества, на основе которых будет производится обучение моделей. Применяется равномерный случайный семплинг с балансировкой классов;

- S_i - обучающие множество, сформированное на основе процедуры семплинга;

- НС (нейронная сеть), ДР (дерево решений), ЛР (лоргистическая регрессия) – базовые модели бинарного классификатора;

- РМ – решающий модуль – реализует алгоритм формирования класса на выходе. Имеет два режима работы – мажоритарный, когда класс определяется простым большинством голосов, и усредняющий – класс определяется как взвешенное среднее «голосов» базовых моделей.

Снижение размерности. Для снижения размерности входных данных в условиях наличия переменных различных типов применение традиционных методов корреляционного и факторного анализа проблематично. Поэтому автором предложено использовать меру значимости, основанную на дивергенции Кульбака-Лейблера, которая показывает степень различия между двумя вероятностными распределениями:

$$D_{\text{КЛ}}(p,q) = \sum_{i=1}^{k} p(\bar{y}) \ln \frac{p(y)}{q(\bar{y})},$$

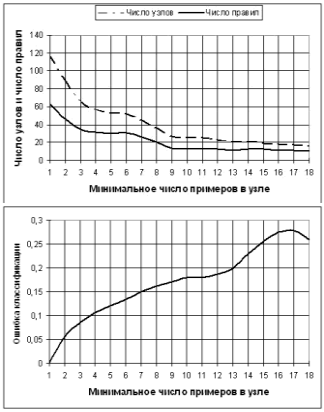
где $p(y)$ и $q(\bar{y})$ - распределения значений бинарной переменной класса. Для этого диапазон изменения каждого признака разбивается на несколько интервалов и вычисляются коэффициенты: $WoE_i = \ln(N_i/N)/(P_i/P)$, где i – индекс интервала, N_i - число не-событий, попавших в интервал, N - общее число не событий в исходном наборе данных, P_i - число событий, попавших в интервал, P - общее число событий. Затем вычисляется величина, называемая информационным индексом

$$IV = \sum_{i=1}^{k} \{(N_i/N - P_i/P) \cdot WoE_i\}$$

и по ее значению выбирается степень значимости признака в соответствии с правилами: $IV < 0{,}02$ отсутствует; $0{,}02 \le IV < 0{,}1$ - низкая; - $0{,}1 \le IV < 0{,}3$ - средняя; - $IV > 0{,}3$ - высокая. В соответствии с данными правилами были выбраны следующие признаки: «Возраст» ($IV = 0{,}29$), «Образование» ($IV = 0{,}1$), «Личный доход» ($IV = 0{,}11$), «Количество ссуд» ($IV = 0{,}29$) «Количество платежей» $IV = 0{,}52$.

Построение НСМ. В качестве базовой архитектуры НС будем использовать персептрон Румельхарта с логистической АФ и алгоритм обучения ОРО. Для оптимизации конфигурации и параметров обучения сети использовалась ошибка классификации, которая для бинарной модели вычисляется как $E = (FP + FN)/(FP + FN + TP + TN)$, где TN –число истинно-отрицательных классификаций, TP – число истинно-положительных классификаций, FN – число ложно-отрицательных примеров классификаций, FP-число ложно-положительных классификаций. Параметры НСМ, обеспечившие наименьшую ошибку обучения: число обучающих примеров – 500, число итераций обучения 2500, число нейронов в скрытом слое L=13, крутизна АФ α=0,4, коэффициент скорости обучения η=0,5 и момент μ=0,45. Ошибка классификации НСМ составила E=0,16.

Построение дерева решений. Полное ДР, построенное на основе алгоритма ID3 содержит 112 узлов и 63 правила. Чтобы сделать ДР более компактным и интерпретируемым, было произведено его упрощение, путем увеличения минимально допустимого числа примеров в узлах. Соответствующие графики и правила результирующего дерева представлены на рис.



а)

Условие			Следствие
Показатель	Знак	Значение	12 Отклик
9.0 Возраст	>=	32,5	
9.0 Личный доход	<	18500	0
9.0 Количество платеж	<	5,5	
9.0 Возраст	>=	32,5	
9.0 Личный доход	<	18500	0
9.0 Количество платеж	>=	5,5	
9.0 Возраст	>=	32,5	1
9.0 Личный доход	>=	18500	
9.0 Возраст	<	32,5	1
ab Образование	=	Среднее специальное	
9.0 Возраст	<	32,5	1
ab Образование	=	Высшее	
9.0 Возраст	<	32,5	1
ab Образование	=	Среднее	
9.0 Возраст	<	32,5	0
ab Образование	=	Неоконченное высшее	
9.0 Возраст	<	32,5	0
ab Образование	=	Неполное среднее	

б)

Рис. 12 – Построение ДР: а) графики зависимости ошибки классификации числа узлов и правил ДР от минимального числа примеров в узле; б) правила ДР.

На рис. 12. видно, что упрощение ДР происходит только до достижения 12-13 примеров в узле, при этом количество правил сокращается до 10, но ошибка возросла до 0,2.

Анализ правил позволил уточнить целевую аудиторию: возраст свыше 32,5 года и личный доход свыше 18 500 руб., а если возраст меньше 32,5 года, то фактором , способствующим отклику клиента является высшее или среднее специальное образование.

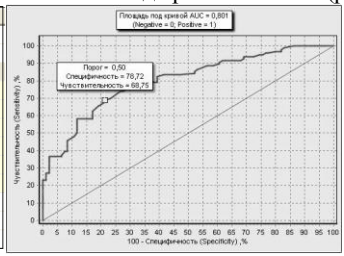
Логистическая регрессия. Модель бинарной классификации на основе логистической регрессии (ЛР), основанная на итеративной подстройке параметров модели, широко применяется в приложениях DM. Параметры модели подстраиваются на каждой итерации в соответствии с правилом:

$$\boldsymbol{\theta}(t+1)=\boldsymbol{\theta}(t)+\alpha\nabla\log L(\boldsymbol{\theta})=\boldsymbol{\theta}(t)+\alpha\sum_{i=1}^{m}\left(y_i-f\left(\boldsymbol{\theta}^T\mathbf{x}_i\right)\right)x_i,\ \alpha>0,$$

где $\boldsymbol{\theta}$ - вектор параметров модели, $\mathbf{x}_i$ - вектор признаков *i*-го примера, α - коэффициент скорости обучения, $y_i\in[0,1]$ - значение бинарной переменной класса для *i*-го примера. Выходом модели является условная вероятность $f(z)=P(y=1|\mathbf{x})$, $\mathbf{x}=(x_1,x_2,..x_n)^T$, где $f(z)=1/(1-e^z)$ - логистичекая функция, $z=\boldsymbol{\theta}^T\mathbf{x}=\theta_1x_1+\theta_2x_2+...+\theta_nx_n$. Таким образом, если задать порог q для вероятности $P(y=1|\mathbf{x})$, такой, что если $P(y=1|\mathbf{x})>q$, то $\hat{y}=1$, в противно случае $\hat{y}=0$.

Построение модели ЛР производилось со следующими параметрами: число итераций *t*=500, коэффициент скорости обучения α=0,5, порог отсечения *q*=0,1. Коэффициенты регрессии представлены на рис. 13, а. Важной задачей при построении модели ЛР является определение порога q , который минимизирует ошибку классификации. Для выбора порога отсечки и оценки качества модели использовался метод кривой ошибок (рис. 13, б, в).

Атрибут	Коэффиц...	Отношение шан...
9.0 <Константа>	-0,58409	
9.0 Возраст	-0,03223	0,96828
ab Образование		
Высшее	0	1
Два и более высших обра...	-0,30998	0,73346
Неоконченное высшее	1,4788	4,3877
Неполное среднее	0,24336	1,2755
Среднее	0,31901	1,3758
Среднее специальное	0,13315	1,1424
9.0 Личный доход	4,0564E-5	1
9.0 Количество ссуд	0,72343	2,0615
9.0 Количество платежей	0,073303	1,0761



а) б) в)

Рис. 13 – Построение модели ЛР: а) коэффициенты регрессии, б) кривая ошибок, в) зависимость ошибки классификации от значения порога отсечки.

Анализ кривой ошибок показал (рис. 13, б), что параметр площади под кривой AUC=0,801, что соответствует высокой точности модели. Значение порога ошибки, обеспечившее минимальную ошибку классификации *q*=0,4. Для трех построенных моделей были получены следующие ошибки классификации: $E_{нс}=0{,}28$, $E_{ЛР}=0{,}35$ и $E_{ДР}=0{,}41$. Ошибка агрегированного классификатора составила $E_{ДР}=0{,}32$.

ОСНОВНЫЕ РЕЗУЛЬТАТЫ И ВЫВОДЫ

1. На основе проведенного в работе обзора и сравнительного анализа инструментальных средств DM и существующих подходов к организации процесса интеллектуальной аналитической обработки данных, была разработана система критериев и классификации аналитических инструментов на основании которых были сделаны выводы и сформулированы

рекомендации по выбору программных средств DM для внедрения аналитических проектов масштаба предприятия;

2. На основе анализа ранее реализованных аналитических проектов в различных проблемных областях были определены основные факторы, влияющие на успешное внедрение аналитических DM-проектов на уровне специалистов, непосредственно интегрированных в процессы управления в социальных и экономических системах, разработана модель для оценки сложности аналитических DM-проектов. Применение данных моделей позволило сократить среднее время разработки и внедрения аналитических проектов сравнимой сложности платформе Deductor на 7%;

3. Разработана концепция сценарного подхода к организации интеллектуальной среды аналитического DM-приложения на основе межотраслевого стандарта организации интеллектуального анализа данных CRISP-DM;

4. Разработаны сценарии построения базовых интеллектуальных моделей на основе нейронных сетей, деревьев решений, карт Кохонена, и интерфейс пользователя для их реализации;

5. Разработана комплексная интеллектуальная модель урожайности зерновых по данным агрохимического обследования почв на основе нейронной сети, дерева решений, карт Кохонена и ассоциативной модели, агрегируемых в ансамбль на основе алгоритма стекинга. Практическое внедрение модели на предприятиях АПК, специализирующихся на выращивании зерновых, позволило снизить среднюю себестоимость продукции на 3,2% и повысить среднюю урожайность на опытных полях на 5,2%;

6. Разработана комплексная модель для анализа клиентской базы кредитной организации на основе ансамбля моделей, основанных на машинном обучении. Практическое внедрение модели позволило повысить процент отклика клиентов на рекламные акции, проводимые на основе директ-маркетинга на 16%.

ОСНОВНЫЕ РЕЗУЛЬТАТЫ ДИССЕРТАЦИОННОЙ РАБОТЫ ОТРАЖЕНЫ В СЛЕДУЮЩИХ ПУБЛИКАЦИЯХ

В изданиях, рекомендованных ВАК РФ.

1. Орешков В.И. Интеллектуальный анализ данных как важнейший инструмент формирования интеллектуального капитала организаций // Креативная экономика. – 2011. – №12. – С. 84-89.

2. Васильев Е.П. Орешков В.И. Моделирование урожайности зерновых с использованием метода совокупности доказательств в рамках концепции точного земледелия // Современные проблемы науки и образования. – 2012 (электронный ресурс).

3. Васильев Е.П. Орешков В.И. Совершенствование процесса принятия управленческих решений в экономике и бизнесе на основе применения интеллектуального анализа данных // Фундаментальные исследования. – 2012. - № 9 вып. 4. – С. 965-971.

4. Орешков В.И. Интеллектуальный анализ данных как современный инструмент поддержки принятия решений в экономике и бизнесе // European Social Science Jornal. – 2012 - No. 9 (том 2) – С. 482 – 490.

5. Е.П. Васильев, В.И. Орешков. Моделирование урожайности на основе данных агрохимического обследования почв с помощью метода ассоциативного анализа.// Вестник РГАТУ. – 2012 - № 4 (16) – С. 8 -13.

6. Е.П. Васильев, В.И. Орешков. Кластеризация данных на основе самоорганизующихся карт признаков в задачах управления в социально-экономических системах. // Вестник РГРТУ. – 2013. - № 3 (вып. 45).

Монографии.

7. Паклин Н.Б., **Орешков В.И.** Бизнес-аналитика: от данных к знаниям (+CD). Изд. 2-е, переработанное и дополненное. - СПб.: Питер, 2010.- 700 с.

Учебные пособия.

8. Васильев Е.П., Орешков В.И.Объектно-ориентированное программирование: реализация экономических задач в среде Delphi. Уч. пособие. – Рязань: РГАТУ, 2011. – 163 с.

Статьи в изданиях, зарегистрированных в Роскомнадзоре.

9. **Орешков В.И**. Интеллектуальный анализ данных как современный инструмент поддержки управленческих решений // Вестник Рязанского гос. агротехнологического университета имени П.А. Костычева. Рязань: РГАТУ. -2011. - №4. - С. 55-59.

10. Васильев Е.П., **Орешков В.И.** Современные аналитические платформы для задач АПК // Вестник Рязанского гос. агротехнологического университета имени П.А. Костычева. Рязань: РГАТУ.- 2011.- № 1.- С.68-75.

Публикации в трудах международных и всероссийских научных и научно-практических конференций.

11. Арустамов А.И., Васильев Е.П., **Орешков В.И**. Интеллектуальные платформы - современный инструмент анализа данных в экономике и бизнесе//Сб. трудов Международной научно-практической конференции «Дни науки», Прага, 2012.

12. Васильев Е.П., **Орешков В.И.** Интеллектуальные системы бизнес-аналитики//Интеграция науки с сельскохозяйственным производством: материалы науч. конф. – Рязань: изд. РГАТУ, 2011 – с. 67-71.

13. Блинкова С.Ю., Васильев Е.П., **Орешков В.И.** Фильтрация данных в интеллектуальных системах бизнес-аналитики//Материалы научно-практической конф. РГАТУ им. П.А. Костычева. Рязань: РГАТУ, 2011 – с. 272 - 277.

14. Васильев Е.П., Воронкина Н.Ю., **Орешков В.И.** Трансформация данных в аналитическом приложении Dedictor Studio// Материалы научно-практической конф. РГАТУ им. П.А. Костычева. Рязань: РГАТУ, 2011 – с. 277 - 282.

15. Васильев Е.П., Гусев Ю.С., **Орешков В.И.** Подавление шумов и сглаживание данных в аналитических системах // Материалы научно-практической конф. РГАТУ им. П.А. Костычева. Рязань: РГАТУ, 2011 – с. 282 - 290.

16. Васильев Е.П., **Орешков В.И.,** Сычева Т.А. Обработка и предобработка данных в задачах АПК// Материалы научно-практической конф. РГАТУ им. П.А. Костычева. Рязань: РГАТУ, 2011 – с. 290 - 296.

17. Васильев Е.П., **Орешков В.И.**, Чумакова Е.Н. Моделирование бизнес-процессов на предприятии АПК в аналитической платформе Deductor// Материалы научно-практической конф. РГАТУ им. П.А. Костычева. Рязань: РГАТУ, 2011 – с. 296 - 304.

18. Васильев Е.П., **Орешков В.И.**, Шаева К.А. Построение модели линейной регрессии в аналитической платформе Deductor// Материалы научно-практической конф. РГАТУ им. П.А. Костычева. Рязань: РГАТУ, 2011 – с. 304 - 310.

Отпечатано в ООО «Полиграф»
390025, г. Рязань, ул. Нахимова, 13.
Тираж 100 экз. Заказ № 133 от 16.05.2013