BaseGroup Labs
ТЕХНОЛОГИИ АНАЛИЗА ДАННЫХ

МЕЖВУЗОВСКАЯ НАУЧНО-ПРАКТИЧЕСКАЯ КОНФЕРЕНЦИЯ «БИЗНЕС-АНАЛИТИКА. ВОПРОСЫ ТЕОРИИ И ПРАКТИКИ. ИСПОЛЬЗОВАНИЕ АНАЛИТИЧЕСКОЙ ПЛАТФОРМЫ DEDUCTOR В ДЕЯТЕЛЬНОСТИ УЧЕБНЫХ ЗАВЕДЕНИЙ»

Deductor

24 июня 2010 года
г. Москва

СБОРНИК МАТЕРИАЛОВ КОНФЕРЕНЦИИ

Рязань 2010

Бизнес-аналитика. Вопросы теории и практики. Использование аналитической платформы Deductor в деятельности учебных заведений: сборник материалов межвуз. науч.-практ. конф. – Рязань: Лаборатория баз данных, 2010. – 155 с.

В сборник материалов научно-практической конференции «Бизнес-аналитика. Вопросы теории и практики. Использование аналитической платформы Deductor в деятельности учебных заведений», проводимой BaseGroup Labs, включены работы преподавателей вузов, активно использующих аналитическую платформу Deductor, из ряда вузов и организаций Москвы, Волгограда, Иваново, Луганска, Нижнего Новгорода, Рязани, Твери, Уфы, Харькова.

Включенные в сборник статьи разделены на две секции: 1) опыт преподавания дисциплин с использованием аналитической платформы Deductor; 2) актуальные задачи бизнес-аналитики и их решение методами Data Mining.

Ответственный редактор сборника материалов конференции: кандидат технических наук **Н.Б. Паклин** (e-mail: education@basegroup.ru)

СОДЕРЖАНИЕ

СЕКЦИЯ «АКТУАЛЬНЫЕ ЗАДАЧИ БИЗНЕС-АНАЛИТИКИ И ИХ РЕШЕНИЕ АЛГОРИТМАМИ DATA MINING»

ПРИВЕТСТВЕННОЕ СЛОВО

24 июня 2010 года на базе института информационных и инновационных технологий НОУ «Международная Академия Бизнеса и Управления» (г. Москва) состоялся уже традиционный съезд вузов-партнеров BaseGroup Labs. В этом году он проходил в формате межвузовской научно-практической конференции *«Бизнес-аналитика. Вопросы теории и практики. Использование аналитической платформы Deductor в деятельности учебных заведений»*.

Это ежегодное мероприятие делается как для существующих преподавателей вузов-партнеров BaseGroup Labs, использующих аналитическую платформу *Deductor* в учебном процессе, так и для новых представителей высших учебных заведений, желающих ознакомиться с образовательной инициативой BaseGroup Labs. К конференции проявляют интерес руководители структурных подразделений вузов, заведующие кафедрами, деканы, преподаватели, желающие применять современные информационно-аналитические системы при обучении студентов – будущих прикладных информатиков, специалистов по информационным системам, экономистов, финансистов, математиков. Впервые очный съезд был успешно проведен в июне 2009 года.

В 2010 году в работе конференции приняли участие 35 участников из 24 вузов России и Украины, в том числе 16 вузов-партнеров BaseGroup Labs. Как обычно, наибольшее число очных участников было из Москвы, но также приехали преподаватели из Твери, Нижнего Новгорода, Иваново и Магнитогорска.


Работа конференции

По тематике присланных докладов и обсуждаемым вопросам можно выделить следующие направления работы конференции.

1. Опыт преподавания дисциплин с использованием аналитической платформы Deductor.

2. Прикладные и исследовательские работы студентов, выполненные на аналитической платформе *Deductor*.
3. Актуальные задачи бизнес-аналитики и их решение алгоритмами Data Mining.
4. Проблемы формирования программ учебных дисциплин, связанных с анализом данных.

Работа конференции началась с приветственного слова руководства Международной Академии Бизнеса и Управления – проректора по учебной работе к.и.н., доцента *Тарасенко Ильи Вадимовича* и директора Института информационных и инновационных технологий МАБИУ к.т.н., доцента *Сергеева Сергея Александровича*.

Тарасенко И.В.

Сергеев С.А.

Конференцию открыл доклад руководителя группы корпоративного обучения BaseGroup Labs к.т.н. *Николая Паклина*. Он рассказал о результатах образовательных инициатив компании за пятилетний период, самые главные из которых – 70 официальных вузов-партнеров, образовательный портал с дистанционными курсами, учебники и книги, выпущенные в центральных издательствах, а также новые формы образовательного взаимодействия – вебинары, онлайн-консультации, конкурс лучших дипломных проектов, выполненных с использованием аналитической платформы *Deductor* и многие другие.

Н.Б. Паклин

Следующий доклад делал директор BaseGroup Labs Алексей Арустамов. Он рассказал о стратегии развития компании на ближайшие годы и роли в ней вузов-партнеров. Большая часть доклада была посвящена эволюции аналитической платформы и ее дальнейшему развитию, а также новой концепции сайта BaseGroup Labs.

Во второй части конференции выступали участники из вузов. Из доклада доцента *Карпузовой Веры Ивановны* (кафедра экономической кибернетики РГАУ-МСХА) можно было узнать об успехах применения аналитической платформы *Deductor* в учебном процессе университета, которая применяется там с 2006 года. За эти годы более 1000 студентов экономических специальностей прошли обучение современным технологиям анализа данных, а также аспиранты, магистранты и слушатели курсов повышения квалификации. Особую роль играет *Deductor* в научной работе студентов; доклады, подготовленные студентами неоднократно отмечались призовыми местами на московских внутривузовских конференциях.

Следующий доклад был сделан профессором МИРЭА, д.т.н. *Болотовой Людмилой Сергеевной*. В ее выступлении затрагивалось много тем: от проблемы отсутствия методологии проектирования хранилищ данных до реальных применений *Deductor* в дипломном проектировании студентов для разработки систем принятия решений по противодействию эпидемиям гриппа (совместно с ГУ НИИ эпидемиологии и микробиологии имени Н.Ф. Гамалеи). Профессор Болотова Л.С. выступила с инициативой создания ассоциации вузов-партнеров с целью их более активного взаимодействия, озвучила пожелания разработчикам аналитической платформы.

В докладе *Потюпкина А.Ю.*, начальника кафедры Военной академии РВСН имени Петра Великого, рассказывалось об особенностях применения методов Data Mining в задачах контроля состояния сложных технических систем, концепции ХД в задачах контроля состояния и перспективах в этой области.

Аспирант МГТУ им. Баумана *Иван Нейский* выступил с докладом о разработанном алгоритме адаптивной кластеризации, который был апробирован для сегментации клиентов брокерского обслуживания и сравнил его эффективность с базовыми алгоритмами k-means и карты Кохонена, для чего использовал аналитическую платформу *Deductor*.

В докладе *Прокопенко Натальи Юрьевны* из ННГАСУ (г. Нижний Новгород) был представлен своеобразный годовой отчет о применении Deductor в учебном процессе специальности «Прикладная информатика», продемонстрированы несколько выпускных квалификационных работ, выполненных студентами по тематике информационно-аналитических систем.

В последнем сообщении *Николая Паклина* было анонсировано новое издание книги «Бизнес-аналитика: от данных к знаниям», которое выходит в печать в июле 2010 года.

В завершении участники конференции пришли к мнению, что по-

добные мероприятия важны и необходимы для вузов, обеспечивают обмен мнениями и опытом в преподавании бизнес-аналитики. Предложено продолжить ежегодные встречи в формате конференции и в следующем, 2011 году снова на базе одного из московских вузов-партнеров.

В ваших руках находится сборник трудов этой конференции. В нем вы найдете печатные версии сделанных очных докладов, а также присланные заочные работы. Мы надеемся, что с каждым годом интерес к нашему мероприятию будет расти и привлекать все больше вузов. До новых встреч!

Компания BaseGroup Labs выражает благодарность руководству Международной Академии Бизнеса и Управления и лично доценту, к. ф.-м. н. *Виктору Викторовичу Татаринову* за помощь в организации и проведении конференции.

ОБРАЗОВАТЕЛЬНЫЕ ИНИЦИАТИВЫ BASEGROUP LABS 2005-2010 гг.

Паклин Н.Б., *к.т.н., BaseGroup Labs*

Нынешний год, 2010, в каком-то смысле знаковый для BaseGroup Labs: прошло пять лет с момента старта образовательных инициатив нашей компании. Поэтому нам открылась хорошая возможность оглянуться назад и проанализировать проделанную работу.

Все начиналось в середине 2005 года, когда стало очевидно – современная обстановка требует начинать изучение информационно-аналитических систем уже на студенческой скамье. Еще в конце 90-х гг. слова и аббревиатуры OLAP, хранилища данных, нейронные сети, кластеризация практически не встречались в учебных программах. Все изменилось за каких-то 5-7 лет. Сегодня информационно-аналитические технологии изучают будущие прикладные информатики, программисты, экономисты. В высшей школе быстро сформировалась потребность на соответствующее программное обеспечение для проведения лабораторных работ. Почувствовав эту тенденцию, наша компания запустила первую инициативу – бесплатно предоставлять комплект лицензий на оборудование компьютерного класса профессиональной версией Deductor вузам, заключившим соответствующее соглашение.

За два следующих года – 2005 и 2006 – нашими вузами-партнерами стали 25 вузов из России, Беларуси и Украины (рис. 1, 2). *Deductor Studio* побеждал тем, что был легким приложением, но вместе с тем имел развитые ETL-средства и основные алгоритмы Data Mining.

Рис. 1. Динамика подключения вузов-партнеров по годам

Рис. 2. Динамика подключения вузов-партнеров по годам

Собственный веб-ресурс BaseGroup Labs к тому моменту насчитывал десятки научно-популярных статей по алгоритмам и технологиям Data Mining, поэтому преподаватель вуза-партнера, потратив определенное время, мог сформировать на их основе конспекты лекций. Для лабораторных работ еще в 2005 году нами был разработан небольшой набор методических материалов по хранилищам данных, OLAP-отчетности и прогнозированию спроса, который поставлялся вместе с данными. Кроме того, каждый преподаватель вуза-партнера приезжал на двухдневный курс обучения в офис BaseGroup Labs. Но к началу 2007 г. становится понятным, что это не решение всех проблем в части методического обеспечения преподавателей вузов-партнеров. Необходим системный курс, объем которого выходил бы за рамки двухдневного тренинга. Кроме того, бесплатная версия *Deductor Lite*, которую использовали вузы, не входящие в партнерскую программу, имела существенное ограничение на объем обрабатываемых записей. Поэтому наша компания принимает несколько решений, существенно повлиявших на дальнейшее развитие образовательной программы.

Во-первых, мы отказались от демо-версии *Deductor Lite*, выпустив вместо нее *Deductor Academic*. В нем отсутствует ограничение на объем обрабатываемых записей, значит, все аналитические алгоритмы становятся доступными при любом числе экспериментальных примеров.

Во-вторых, закупается полноценная система дистанционного обучения (СДО) и начинается разработка системного курса «Корпоративные аналитические системы», который уже в сентябре 2007 года придет на смену очным двухдневным тренингам преподавателей вузов-партнеров.

Этот дистанционный курс, разбитый на 12 разделов, содержит более 90 конспектов, 15 практикумов, 18 индивидуальных бизнес-задач и свыше 600 аттестационных вопросов по всем разделам. Его объем со-

ставляет более 200 академических часов. При его разработке использовался весь накопленный практический опыт аналитиков BaseGroup Labs при внедрении реальных проектов, а при написании теоретических конспектов мы ориентировались на лучшие западные книги и учебники (к сожалению, в учебниках по бизнес-аналитике и Data Mining Россия отстала от Запада на несколько лет).

Появление дистанционного курса и его реализация на современных e-learning технологиях позволило нам ввести систему сертификации аналитиков Deductor (рис. 3). К середине 2010 г. свыше 40 преподавателей вузов-партнеров прошло обучение либо продолжают учиться на образовательном портале BaseGroup Labs, 18 из них получили статус сертифицированного аналитика. К сожалению, среди преподавателей вузов не удается достичь 100% сертификации, но эта проблема заслуживает отдельного разговора.

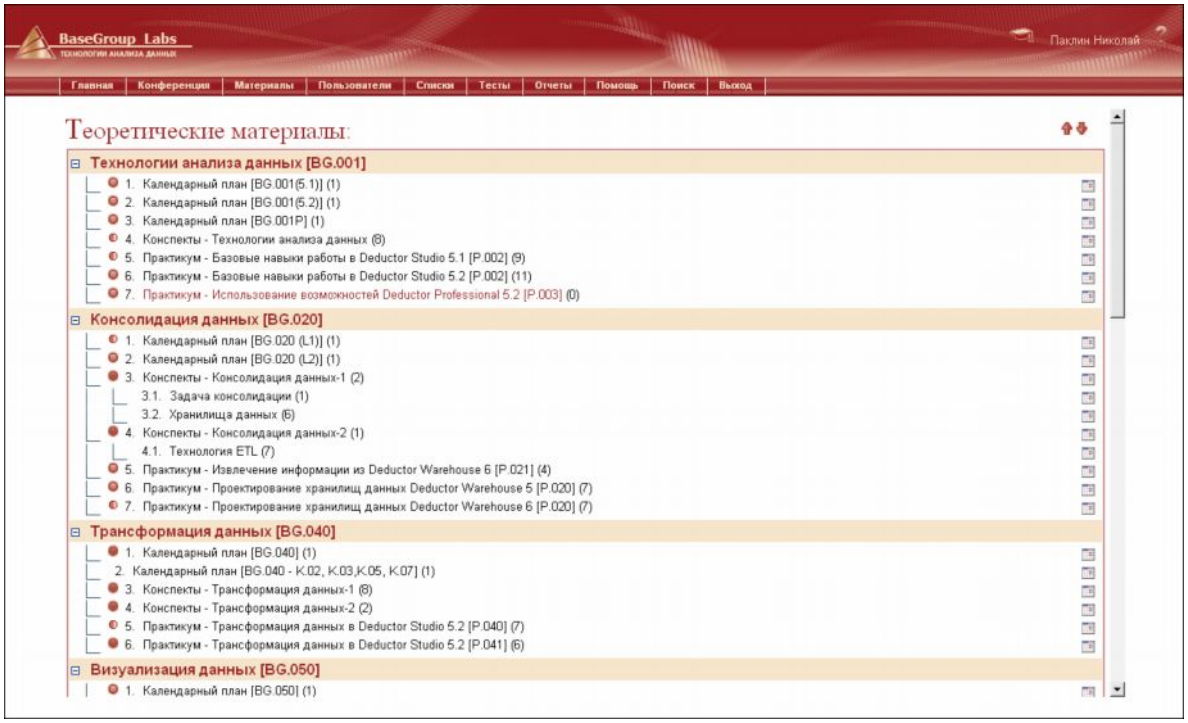


Рис. 3. Система дистанционного обучения BaseGroup Labs

В 2009 году образовательные инициативы продолжаются. Выходят в свет книги в центральных издательствах: «Практикум по анализу данных на компьютере» (И. Кацко, Н. Паклин, учебное пособие с грифом УМО, издательство «КолосС»), «Бизнес-аналитика: от данных к знаниям» (Н. Паклин, В. Орешков, издательство «Питер»). В основу последней книги лег дистанционный курс компании, и она имела большой успех на российском книжном рынке, оказалась востребованной преподавателями и студентами вузов, тираж в 1500 экземпляров был раскуплен в считан-

ные месяцы. Через 1,5 года, в июле 2010 года, выйдет 2-е издание книги, переработанное и дополненное.

В этом же году совместно с НОУ «ИЭФ «Синергия» при РЭА им. Г.В. Плеханова открываются совместные программы повышения квалификации «Управление корпоративными аналитическими системами» с выдачей свидетельства государственного образца. Дистанционную составляющую программ обеспечивает BaseGroup Labs. На образовательном портале возможность проходить обучение получают частные лица. В июне 2009 г. на базе «Синергии» проходит первый семинар для вузов-партнеров.

Развитие новых форм коммуникаций и рост пропускной способности интернет-каналов позволило нам в 2010 году освоить новые формы взаимодействия – вебинары и онлайн-консультации тьюторов образовательного портала. Мы будем развивать эти формы и дальше.

Число официальных вузов-партнеров к середине 2010 года достигло числа 70. По нашим приблизительным оценкам, еще около 100 вузов используют аналитическую платформу *Deductor* в учебном процессе. Кроме того, нам удалось довольно точно подсчитать количество инсталляций *Deductor Academic* благодаря встроенной странице обновлений, открываемой при старте *Deductor*. Выяснилось, что насчитывается свыше 5 100 пользователей *Deductor Academic 5.2*. Можно смело предположить, что на сегодня это самая популярная свободно распространяемая аналитическая платформа в русскоязычном Интернете!

Желание нашей компании стимулировать вузы становиться нашими партнерами можно объяснить анонсирование двух ежегодных конкурсов, которые будут проходить регулярно: «Лучший вуз-партнер» и «Конкурс студенческих дипломных работ, выполненных с использованием аналитической платформы *Deductor*». Итоги подводятся каждый сентябрь независимым жюри, и все победители будут награждены ценными призами: оригинальные кубки, фирменные грамоты, лицензионные ключи *Deductor Studio Professional*, книги. А первая номинация предназначена для поощрения учебно-методической и научно-исследовательской работы вузов-партнеров с применением *Deductor*.

На первый конкурс дипломных работ было прислано немного заявок, однако все работы интересные. Это и решение традиционных бизнес-задач («Разработка автоматизированной информационной системы принятия «инвестиционных решений на базе АП Deductor», «Применение современных информационных технологий и интеллектуальных методов анализа в задаче оценки недвижимости», ННГАСУ, Нижний Новогород, науч. рук. доцент *Прокопенко Н.Ю.*, «Модернизация системы информационного обеспечения кредитования физических

лиц коммерческого банка», ИГЭТУ, Иваново, науч. рук. доцент *Баллод Б.А.*), так и технические приложения («Система биометрической аутентификации пользователей по распознаванию движений мыши», КГТУ им. Туполева, Казань, науч. рук. доцент *Катасев А.С.*).

Конечно, не все 70 вузов-партнеров являются активными, таких не больше 15, хотя во многих случаях проблема в том, что наша компания предоставляет вузам мало способов для обратной связи. Тем не менее, нельзя не упомянуть инициативы вузов-партнеров, которые имеют законченный результат. Это совместная работа профессора *И.А. Кацко* и доцента *Н.Б.Паклина* по созданию учебно-методического пособия «Практикум по анализу данных на компьютере»; методическая разработка по работе с *Deductor* для экономистов коллектива авторов из РГАУ-МСХА во главе с доцентом *Карпузовой В.И.* с оригинальными примерами из области сельского хозяйства; научная работа кафедры математических методов в социологии Алтайского государственного университета в области сегментации рынка труда, в ходе которой использовалась АП Deductor, а позже – опубликована монография (руководитель – доцент *Мальцева А.В.*).

Проявляют активность вузы-партнеры Украины. Так, *доцент Хомич С.В.* (г. Ровно, РГГУ) на основе методических материалов BaseGroup Labs разработал сайт для студентов с презентациями, лекциями и практикумами на украинском языке по аналитической платформе *Deductor* и бизнес-аналитике.

Более подробно ознакомиться с деятельностью вузов-партнеров, а также узнать о том, как можно стать вузом-партнером, можно на сайте нашего образовательного портала по адресу: edu.basegroup.ru.

Компания BaseGroup Labs всегда уделяла большое внимание работе с вузами. Интерес к нашим разработкам, аналитической платформе, книгам, конференциям, конкурсам убедительно говорит о том, что вместе мы вносим большой вклад в подготовку грамотных и разбирающихся в современной бизнес-аналитике специалистов.

СЕКЦИЯ
«ОПЫТ ПРЕПОДАВАНИЯ ДИСЦИПЛИН С ИСПОЛЬЗОВАНИЕМ АНАЛИТИЧЕСКОЙ ПЛАТФОРМЫ DEDUCTOR»

О ПОДХОДЕ К ПРЕПОДАВАНИЮ СИСТЕМ БИЗНЕС-АНАЛИЗА В ЭКОНОМИЧЕСКОМ ВУЗЕ

Завьялова Н.Б., *доцент Российской экономической академии им. Г.В. Плеханова, г. Москва*
Дьяконова Л.П., *доцент Российской экономической академии им. Г.В. Плеханова, г. Москва*

В настоящее время в комплексе современных инструментальных средств, обеспечивающих поддержку бизнеса, значительную роль играют аналитические инструментальные средства. Развитие инфраструктуры компании и спектр решаемых аналитических задач зависит от уровня зрелости организации, ее стратегических и тактических целей, а также степени подготовки специалистов и их мотивации для использования инновационных технологий.

В РЭА им. Г. В. Плеханова изучение систем бизнес-анализа (BI) включено в программы магистров и бакалавров по направлениям «Экономика» и «Менеджмент» в рамках таких дисциплин как «Информационные технологии в управлении организацией», «Компьютерные технологии в бизнес-исследованиях» и др. Целью этих дисциплин (или отдельных модулей) является формирование у студентов целостного представления о подходах к решению задач бизнес-анализа, принципах проектирования, внедрения и эксплуатации информационных аналитических систем на стратегическом, тактическом и оперативном уровнях управления экономическими объектами.

С методической точки зрения процесс преподавания в большинстве случаев реализуется в соответствии с алгоритмом, представленным на рис. 1.

Этапы 1-3 позволяют направить исследовательскую работу студентов на детальное изучение предметной области и проблем, подлежащих автоматизации средствами информационных технологий и систем.

Этап 4 «Выбор инструментальных средств» является чрезвычайно важным элементом исследовательского проекта, поскольку успех внедрения и модификации информационных систем во многом определяется соответствием ожиданий заказчиков реальным функциональным возможностям системы. На данном этапе осуществляется не только всесторонний анализ функциональных возможностей аналитических систем и приложений, но и оценка затрат на внедрение и эксплуатацию, адек-

ватность лицензионной политики поставщика решения, доступность и простота эксплуатации, наличие специалистов и развитых средств поддержки клиентов.

Рис. 1. Пример алгоритма реализации исследовательской работы

Слушатели должны продемонстрировать свое знание основных классов аналитических систем, провести сравнительный анализ и осуществить мотивированный выбор применяемых аналитических инструментов.

Спектр инструментальных средств бизнес-анализа чрезвычайно велик, поэтому оценку и фиксацию результатов сравнительного анализа слушателям предлагается выполнить средствами функционального моделирования (программные продукты *ARIS, BPwin/ AllFusion и др)* в рамках модели, первый иерархический уровень которой представлен на рис. 2. Такой подход позволяет не только структурировать процесс выбора инструментальных средств, но и также сформировать отчет в HTML-формате с последующей его публикацией на сайте. Примером является декомпозиция работы А2 с последующим разбиением на подмодели ис-

следований по каждому программному продукту.

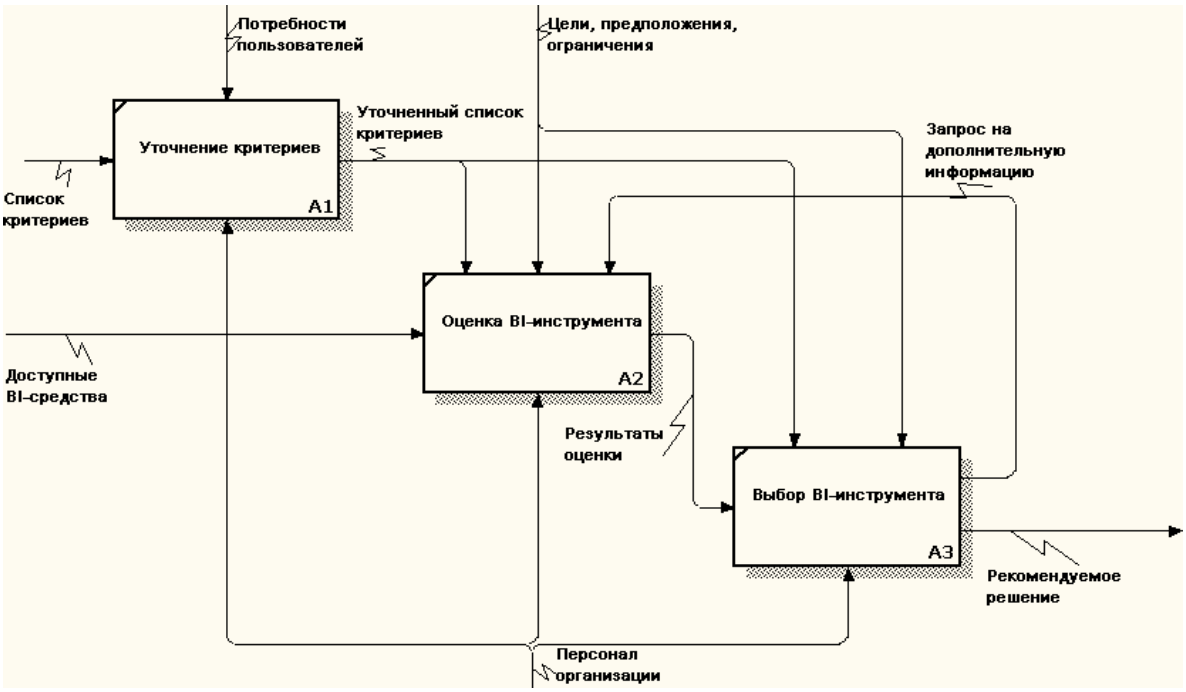


Рис. 2 . Модель оценки процесса выбора BI-инструментария

Этапы 5-6 представляют собой основу проекта и содержат аналитические исследования бизнес-ситуаций в компании.

Примером исследовательских работ может служить курсовой проект «Прогнозирование продаж продукции предприятия хлебопекарной промышленности», в котором перед слушателями ставится задача анализа временного ряда продаж категории «Массовые сорта» на интервале устойчивого заказа на производство продукции (60 дней), изучения возможностей инструментальных средств и анализ качества реализованных моделей.

Следует отметить, что прогнозирование ключевых показателей маркетингового планирования является наиболее сложным видом исследований, так как спрос и продажи зависят от сезонности, динамики развития бизнеса, конкурентной среды, ценовой политики, специфических особенностей продукции.

Экспресс-анализ данных с помощью хорошо изученных инструментальных средств, например, MS Excel (рис. 3), позволяет исследователям сосредоточиться на изучении предметной области, получить предварительные результаты и сформулировать предварительные выводы. В данном случае очевидно, что массовые сорта продукции играют доминирующую роль в объемах продаж предприятия; снижение объемов продаж всех категорий продукции наблюдается в субботние и выходные дни; аномально низкие значения данных выявлены в первую неделю года, что

обусловлено значительным снижением спроса в праздничные дни; нулевые данные или пропуски данных отсутствуют для массовых сортов продукции, однако наблюдаются в других исследуемых рядах.

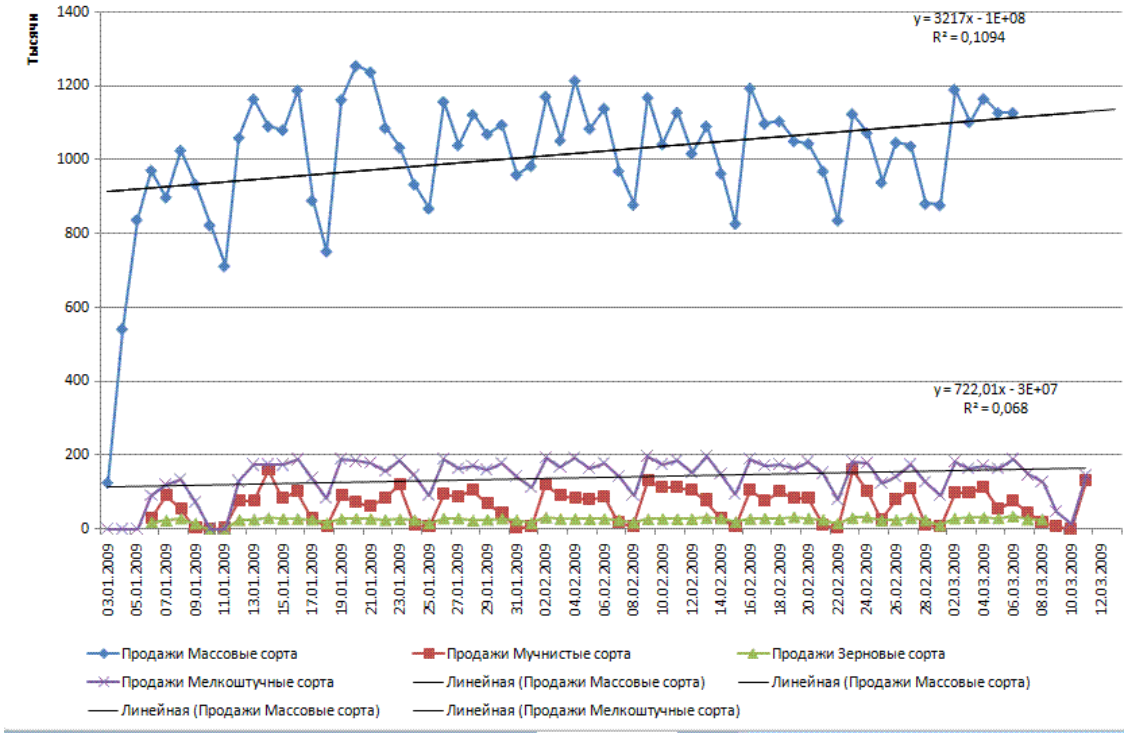


Рис. 3. Продажи продукции

На стадии углубленного анализа слушателям предстоит выполнить решение задачи прогнозирования, проведя исследования средствами различных информационных аналитических систем. В рамках проекта «Прогнозирование продаж продукции предприятия хлебопекарной промышленности» изучались возможности построения прогнозных моделей продаж средствами информационно-аналитической системы *Marketing Analytic* (компания Курс), аналитической платформы *Deductor* (компания BaseGroup Labs) и статистического пакета *Statistica* (компания StatSoft).

При выполнении этапа 5 «Разработка и корректировка модели бизнес-исследования» слушатели демонстрируют владение инструментальными средствами и методикой анализа.

Так, в рамках рассматриваемого проекта был выполнен прогноз продаж методом авторегрессии без выделения сезонности путем настройки параметров ретропрогноза: порядка авторегрессии и собственных значений. Исследования выполнялись средствами *Marketing Analytic* путем настройки следующих параметров: порядок авторегрессии – 9, число собственных значений – 7, число точек – 10, выделен линейный тренд. Результаты исследования (рис. 4) показали удовлетворительное качество модели (ошибка прогноза не превышает 10%).

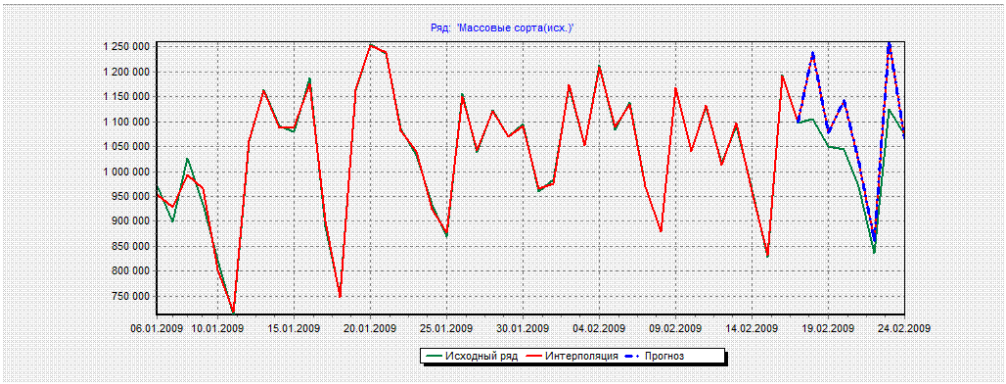


Рис. 4. Результаты прогноза продаж массовых сортов продукции

Аналитическая платформа *Deductor* предоставляет более широкие возможности для предварительной обработки и анализа данных. К рассматриваемому временному ряду были применены операции сглаживания и фильтрации, реализуемые в обработчике «Парциальная обработка», что позволило снизить высокочастотные колебания внутри недельного цикла продаж.

Построение моделей прогноза продаж средствами аналитической платформы *Deductor* осуществлялось по алгоритму, реализованному в виде сценария, представленного на рис. 5.

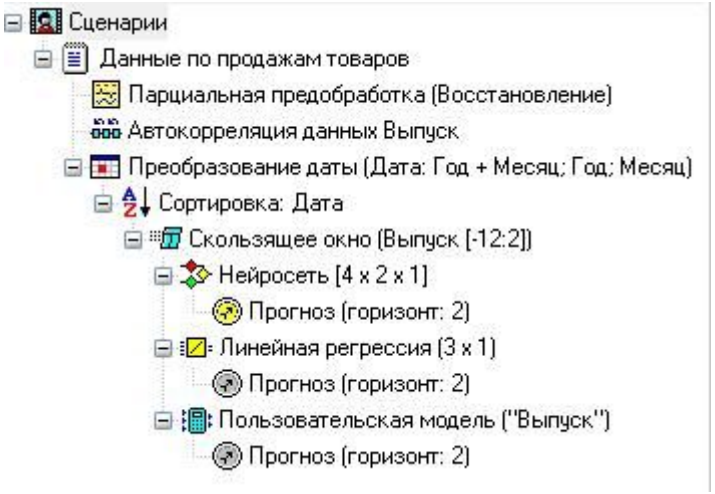


Рис. 5. Сценарий прогнозирования продаж

Алгоритм построения нелинейной модели на основе искусственной нейросети сводился к определению архитектуры сети и запуска процесса обучения, позволяющего определить оптимальные значения весовых коэффициентов. При разбиении исходного набора данных на подмножества 95% выделялось для обучающего подмножества и 5% - для тестового. В качестве активационной функции была выбрана сигмоида, а условием прекращения обучения является ошибка меньше 0,05 по дос-

тижении эпохи 10000.

Анализ диаграммы рассеяния показал, что разброс между эталонными значениями выходного поля и значениями, рассчитанными моделью, достаточно невелик, а ошибка прогноза не превысила 8 % (рис. 6, рис. 7).

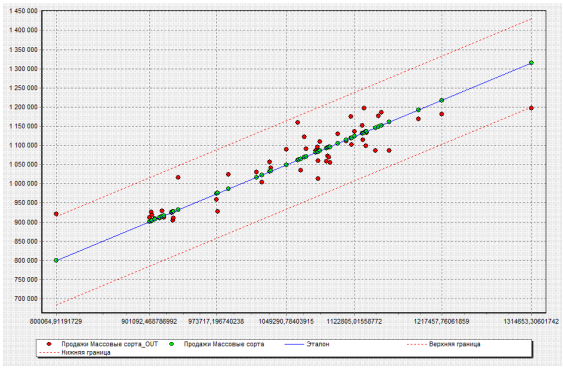


Рис 6. Диаграмма рассеяния

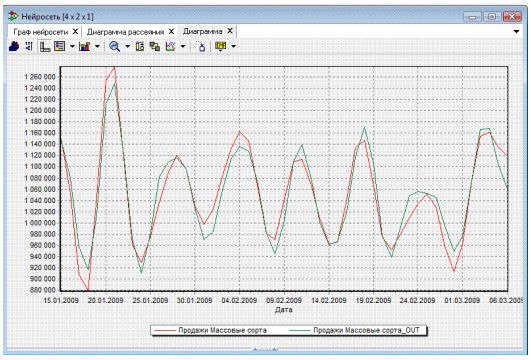


Рис. 7. Исходный и аппроксимирующий ряды

Реализованная по сценарию, представленному на рис. 5, линейная модель показала удовлетворительную качество модели на диаграмме рассеяния (рис. 8, рис. 9) и низкую ошибку прогноза (2%).

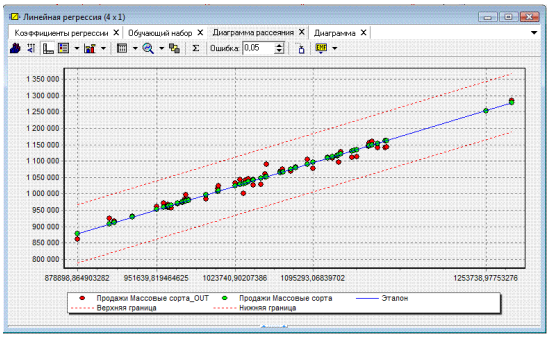


Рис. 8. Диаграмма рассеяния

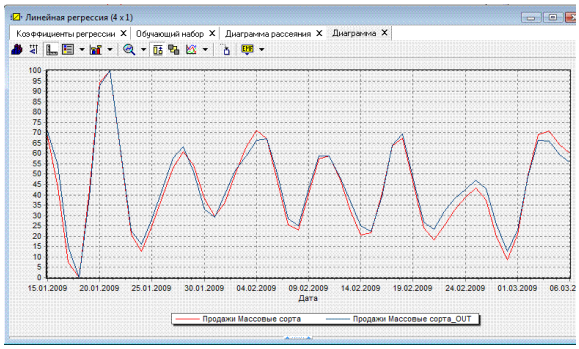


Рис. 9. Исходный и аппроксимирующий ряды

Пакет *Statistica* предоставляет широкие возможности по анализу и прогнозированию временных рядов. Модуль *Временные ряды и прогнозирование* пакета *Statistica* содержит большой набор процедур по обнаружению систематических компонент ряда, построению автокорреляционных функций, визуализации данных.

Для достижения целей исследовательской работы и решения задачи прогнозирования рассматривались методы экспоненциального сглаживания и АРПСС. Практика применения метода АРПСС подтвердила его мощность и гибкость, однако потребовала от слушателей серьезных затрат времени на его освоение.

На начальной стадии исследования с применением метода АРПСС

была проведена идентификация порядка модели – определение числа параметров авторегрессии (p) и скользящего среднего (q). Значения параметров моделей подбирались по первым 50 наблюдениям ряда, прогноз (ретропрогноз) проводился на следующие 10 дней. В результате анализа графиков (корелограмм) АКФ и ЧАКФ был сделан выбор модели: АРПСС (0,1,1)(0,1,1), сезонный лаг 7.

Анализ остатков показал, что они некоррелированы и нормально распределены.

На рис. 10 показан график исходного ряда и ретропрогноза на 10 дней для данной модели.

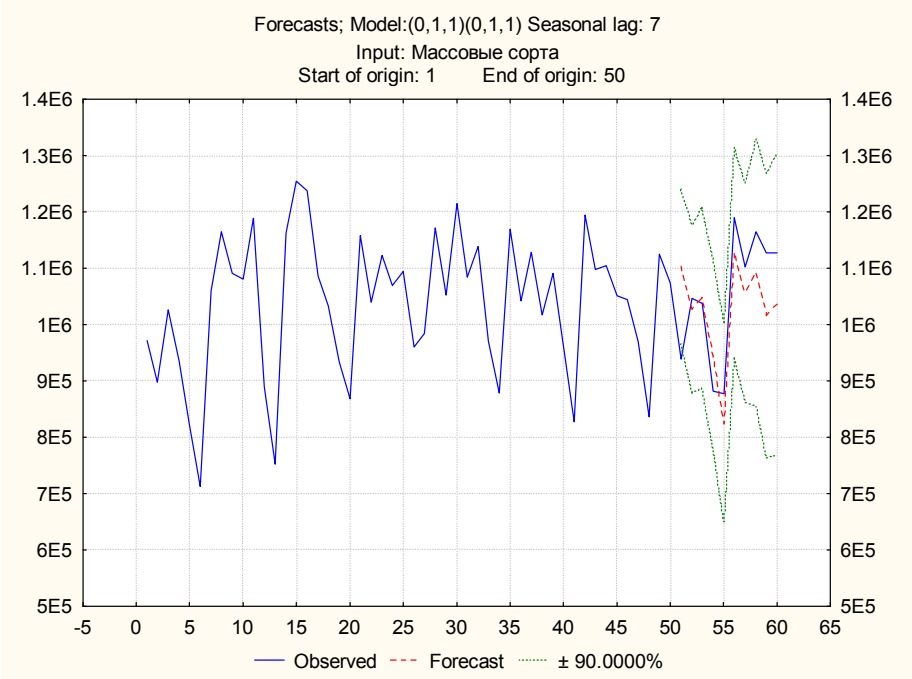


Рис. 10. Прогнозирование методом АРПСС

Результатом работы слушателей должны стать выводы в отношении качества прогнозов, полученных несколькими методами и в различных системах, и формирование таблицы-заключения по возможности решения задачи прогнозирования продаж.

Следует отметить, что значительная часть специалистов, занимающихся маркетингом, имеет недостаточный уровень подготовки в области статистической обработки наблюдений. В связи с этим возникает потребность в программах, которые по своему научному уровню были бы достаточно совершенными, а по интерфейсу и технологии обработки информации – простыми. В то же время немаловажным требованием является точность прогнозов, обеспечивающая качество принятия решений.

Комплексный подход к изучению технологий бизнес-анализа с ис-

пользованием различных инструментальных средств и методов позволяет расширить кругозор слушателей и, что самое главное, привить навыки исследовательской работы и принятия решений в области автоматизации бизнеса.

Литература

1. Паклин Н.Б., Орешков В.И. Бизнес-аналитика: от данных к знаниям (+CD). – Спб.: Питер, 2009. ─ 624 с.: ил.
2. http://www.basegroup.ru/ – информационный ресурс компании BaseGroup Labs.

ПОДГОТОВКА СТУДЕНТОВ-БАКАЛАВРОВ К РЕШЕНИЮ ЗАДАЧ МАЛОГО БИЗНЕСА МЕТОДАМИ ИННОВАЦИОННЫХ ТЕХНОЛОГИЙ

Кудряшова Э.Е., *доцент, профессор РАЕ, г. Волгоград*

При вхождении России в систему Европейского образования в соответствии с Болонской конвенцией структура дисциплин на базе бакалавриата в вузах не содержит ряд тем, необходимых современному специалисту с высшим образованием. Если рассматривать выпускников-бакалавров как специалистов, подготовленных к трудоустройству на предприятиях малого бизнеса на современном уровне, то недостаточными являются знания инновационных технологий в области искусственного интеллекта, автоматизации задач бизнес-планирования и т.д.

В условиях сложной экономической ситуации становится актуальным обеспечение специалистов предприятий малого бизнеса автоматизированным инструментарием оперативного принятия управленческих решений. Оперативное принятие управленческих решений, адекватных изменяющимся условиям экономической и социальной среды, позволит гибко управлять хозяйственной деятельностью предприятия малого бизнеса.

В социуме присутствует такая важная общая характеристика, как самоорганизация – одна из форм синергизма. Обобщающий взгляд, характерный для синергетики, обладает большой эвристической ценностью при анализе таких явлений, как «экономика, основанная на знаниях» (knowedge-based economy, инновационная экономика), экономические катастрофы и ряд других. Основой современной «новой экономикки» представляется инновационный взрыв в сфере информационных тех-

нологий (компьютеры, программное обеспечение, телекоммуникации и Internet). Использование новых информационных технологий: методов искусственного интеллекта, компьютерных средств когнитивного моделирования и т.д., открывает новые возможности специалистам предприятия малого бизнеса. Включение представления знаний в автоматизированную систему искусственного интеллекта рассматривается во взаимосвязи с качественными и количественными параметрами когнитивной модели с позиций синергетики.

Для подготовки студентов, обучающихся в структуре бакалавриата, к решению современными методами практических задач малого бизнеса возможно создание неформального межвузовского Студенческого научно-инновационного центра. Студенты, заинтересованные в углублении знаний, могут заниматься на курсах, раскрывающих следующие темы.

Экономика, основанная на знаниях. Рассматриваются методы извлечения знаний на базе интеллектуальных моделей. Демонстрируются интеллектуальные возможности аналитической платформы Deductor при обработке экономических данных.

Онтологическое моделирование при проектировании автоматизированных систем. Разрабатываются онтологические модели на примерах задач предприятий малого бизнеса.

Автоматизация разработки бизнес-процессов на примерах задач предприятий малого бизнеса. Моделирование бизнес-процессов на основе SADT-методики.

Инновации и риски в деятельности предприятия малого бизнеса. Методики стратегического анализа (GAP-анализ, матрица Ансоффа, SWOT-анализ, PEST-анализ, BSC-анализ) в применении к задачам малого бизнеса.

«Качественное» моделирование. Рассматривается когнитивное моделирование деятельности предприятия малого бизнеса в условиях нечёткой многосвязной среды.

Литература

1. Кудряшова Э.Е. Методы и модели проектирования информационных систем / Монография. – М.: «Академия естествознания», 2009. – 128 с.

ОПЫТ ИСПОЛЬЗОВАНИЯ АНАЛИТИЧЕСКОЙ ПЛАТФОРМЫ DEDUCTOR ПРИ ПОДГОТОВКЕ СПЕЦИАЛИСТОВ ДЛЯ ЭКОНОМИКИ, ОСНОВАННОЙ НА ЗНАНИЯХ

Настащук Н.А., *доцент Омского государственного педагогического университета, г. Омск*

Перспективным направлением информатизации сферы экономики является профессиональная деятельность по управлению знаниями, которая основана на применении средств интеллектуальных информационных технологий (ИИТ). В связи с этим в качестве одной из приоритетных целей высшего экономического образования определяется подготовка современного экономиста к профессиональной деятельности по управлению знаниями в экономике и совершенствованию механизмов их воплощения в инновации (Н.И. Лапин, А.М. Лялин, Б.З. Мильнер и др.), а также реализация модели образования для экономики, основанной на знаниях (Я.И. Кузьминов) [5]. Кроме того, решение задачи управления знаниями потребовало внести заметные изменения в организационные структуры финансовых и промышленных компаний, результатом которых стало появление новых должностей: директор по управлению знаниями, вице-президент по управлению интеллектуальным капиталом, менеджер по интеллектуальным активам, брокер знаний, экспедиторы решений и др. [2].

Управление знаниями рассматривается как совокупность процессов, управляющих созданием, распространением, обработкой и использованием знаний в рамках организации [1]. Технологической основой систем управления знаниями являются хранилища данных. Анализ информации в хранилищах данных базируется на технологиях интеллектуального анализа данных, целью которого является извлечение знаний из накопленных данных за некоторый промежуток времени.

Таким образом, появляются новые технологии организации, хранения и обработки экономической информации. Примером таких технологий являются так называемые *Business Intelligence средства*, предоставляющие конечному пользователю возможности доступа и последующего анализа прикладных структурированных данных с целью прогнозирования и принятия решений в сфере экономики и бизнеса [4].

К средствам *Business Intelligence* относятся хранилища данных *Data Warehouse*, генераторы отчетов и средства аналитической обработ-

ки OLAP, а также средства поиска закономерностей – *Data Mining*. *Business Intelligence средства* (или искусство преобразовывать данные в знания) являются одним из аспектов управления знаниями. Вышеуказанные средства в полной мере реализует аналитическая платформа *Deductor*. Хотя, академическая данной системы имеет некоторые ограничения, но это никак не отражается на возможности полноценного изучения ее функционала и, соответственно, подготовке будущих специалистов для экономики, основанной на знаниях.

Обучение использованию аналитической платформы Deductor о для решении предстоящих задач в области управления знаниями рекомендуется проводить при подготовке как студентов информационно-экономического профиля (специальности «Прикладная информатика в экономике», «Прикладная информатика в менеджменте» и направления подготовки «Прикладная информатика», «Бизнес-информатика»), так и экономических специальностей (например, «Бухгалтерский учет, анализ и удит», «Финансы и кредит» и др.).

Определим обобщенную схему решения задач в экономических исследованиях средствами *Business Intelligence*, в частности *Deductor*, следуя логике системного подхода, который определяет методологическую основу современной информационной деятельности специалиста и принятия решений, характерных для управления знаниями [3]:

1. провести анализ предметной области, к которой принадлежит хранилище данных, и очертить круг проблем, характерных для данной предметной области;
2. определить цель или сформулировать гипотезу поставленной задачи;
3. выявить набор существенных атрибутов, которые наиболее точно и полно характеризуют цель поставленной задачи;
4. выбрать необходимое *Business Intelligence* средство и построить модель решения задачи;
5. на основе полученных результатов, в том числе, графических (диаграммы, графики и т.д.), сформулировать выявленные закономерности, поведение фирмы на рынке и др.

Необходимо отметить, что в ходе проведения диссертационного исследования [5], посвященного исследованию проблемы обучения ИИТ и определению возможных путей и средств развития учебно-познавательной компетенции при данном обучении в рамках высшего экономического образования, была использована аналитическая платформа *Deductor* для реализации технологии *Business Intelligence*. Данная технология базируется на интеллектуальном анализе данных, который относится к машинному обучению – одному из актуальных направлений

научной области «Искусственный интеллект». При этом построена и теоретически обоснована модель обучения ИИТ для развития данной компетенции, разработана и реализована методика обучения ИИТ, основанная на использовании учебных эвристических задач с экономическим содержанием.

К эвристическим относят задачи, направленные на постановку проблемы и реализацию ее решения на тактическом и стратегическом уровнях управления предприятием в условиях быстро меняющейся рыночной экономики и внедрением инноваций, зачастую в условиях неопределенности (наличие неполной информации и нечетких исходных данных). Эвристическая задача характеризуется тем, что для нее не известен формальный математический алгоритм, способ решения, а также наличие качественных суждений. При этом познавательная деятельность будущих экономистов сводится к формализации неопределенности и неточности экономической информации, содержащейся в задаче, что требует умений системного видения предмета решаемой задачи [5].

Рассмотрим конкретный пример решения учебной эвристической задачи с экономическим содержанием, которую можно предложить студентам на одной из лабораторных работ в рамках дисциплины «Интеллектуальные информационные системы» (специальности «Прикладная информатика в менеджменте» и «Прикладная информатика в экономике»). Данная задача решается с помощью аналитической платформы Deductor.

Задача. Требуется выявить группы риска заемщиков компании N на основе применения технологии интеллектуального анализа данных (технология *Data Mining*). Для этого следует ответить на вопросы:

- Каков тип кредитоспособных заемщиков?
- Каков тип некредитоспособных заемщиков, которым отказали в выдаче кредита?
- Какие факторы оказывают наибольшее и наименьшее влияние на необходимость отказа в кредите?
- Какие факторы оказывают наибольшее и наименьшее влияние на выдачу кредита?
- Какова стратегия привлечения заемщиков с другими профилями?

Методические рекомендации к решению задачи.

На первом, втором и третьем шагах решения задачи студенту необходимо сформулировать несколько гипотез, исходя из проблематики заданной предметной области. Следующим шагом является выявление существенных и менее существенных факторов, характерных для выдвинутого предположения. Результатом данного шага должен быть список с

описанием всех факторов, которые рекомендуется оформить в виде таблицы (табл. 1).

Таблица 1

Оценка значимости факторов

Фактор	**Экспертная оценка значимости (1 - 100)**
Среднегодовой доход	100
Сумма кредита	90
Семейное положение	80
Пол	45
Возраст	75
Профессия	62
Срок проживания в городе	20
Образование	70
Автомобиль	90

После подготовки таблицы с описанием факторов нужно экспертно оценить значимость каждого из факторов, используя метод экспертных оценок, который способствует снятию неопределенности и выявлению более существенных и менее существенных факторов.

На четвертом шаге необходимо выбрать соответствующую модель для реализации компьютерного интеллектуального анализа информации. На основе полученного результата значимость факторов может меняться и, соответственно, корректируется или отвергается одна из гипотез. В данном случае выбирается такой метод *Data Mining* как дерево решений.

На пятом шаге в результате проведенного компьютерного анализа выяснилось, что кредитоспособными являются следующие лица: 1) мужчины и женщины, берущие сумму кредита меньше 24 тыс. руб., срок проживания в данной местности меньше 11 лет и среднегодовой доход составляет 160 – 180 тыс. руб. в год; 2) мужчины в возрасте 28 – 48 лет с высшим или средним образованием, наличие автомобиля, срок проживания в данной местности больше 11 лет. Далее студент, руководствуясь извлеченными знаниями из базы данных, отвечает на поставленные вопросы задачи и формулирует окончательное решение задачи, используя метод конструирования правил.

Например:

Если *физическое лицо обладает <перечисление выявленных значимых условий>,* ***То*** *это физическое лицо является кредитоспособным.*

Если *физическое лицо обладает <перечисление выявленных значимых условий>,* ***То*** *это физическое лицо не является кредитоспособным.*

Затем студентам рекомендуется продолжить решение данной задачи в следующем творческом направлении: разработайте тактику выдачи кредита населению согласно выявленным значимым и менее значимым факторам, влияющим на кредитоспособность.

Методика обучения решению задач, ориентированных на управление знаниями, посредством применения аналитической платформы Deductor внедрена и продолжает совершенствоваться при подготовке студентов по следующим специальностям: «Прикладная информатика в менеджменте» на факультете информатики Омского государственного педагогического университета, а также «Бухгалтерский учет, анализ и аудит», «Управление персоналом», «Прикладная информатика в экономике» и направления подготовки «Прикладная информатика» Омского экономического института.

Таким образом, аналитическая платформа Deductor играет значимую роль при подготовке специалистов для экономики, основанной на знаниях, при этом наглядно демонстрирует функциональные возможности *Business Intelligence* средств, которые составляют технологическую основу управления знаниями.

Литература

1. Гаврилова Т.А. Логико-лингвистическое управление как введение в управление знаниями // Новости искусственного интеллекта. 2002. – № 6. – С. 36–40.
2. Мильнер Б. З. Управление знаниями в современной экономике / Б.З. Мильнер // Проблемы теории и практики управления. – 2006. – № 9. – С. 8 – 13.
3. Настащук Н. А. Системный подход к использованию Business Intelligence средств в экономических исследованиях / Н. А. Настащук // Информационная среда вуза: материалы XIV Междунар. науч.-техн. конф. – Иваново, 2007. – С. 302-305.
4. Настащук Н. А. Некоторые аспекты использования Business Intelligence средств в профессиональной подготовке студентов специальности "Прикладная информатика в экономике" / Н. А. Настащук// Научно-практический журнал "Экономика, статистика и информатика. Вестник УМО". – 2008. – № 2. – С. 69-71.
5. Развитие учебно-познавательной компетенции у будущих экономистов в процессе обучения интеллектуальным информационным технологиям : автореф. дис. на соиск. учен. степ. канд. пед. наук / Н. А. Настащук; Омск. гос. пед. ун-т. – Омск, 2009. – 21 с.

ИСПОЛЬЗОВАНИЕ АНАЛИТИЧЕСКОЙ ПЛАТФОРМЫ DEDUCTOR ПРИ ПОДГОТОВКЕ СПЕЦИАЛИСТОВ ЭКОНОМИЧЕСКОГО ПРОФИЛЯ

***Карпузова В.И.**, доцент, **Скрипченко Э.Н.**, профессор, **Чернышева К.В.**, доцент, Российский государственный аграрный университет – МСХА имени К.А.Тимирязева, г. Москва*

В Российском государственном аграрном университете – МСХА имени К.А. Тимирязева большое внимание уделяется подготовке высококвалифицированных кадров, способных решать задачи современной экономики с использованием новейших информационных систем и технологий.

В зависимости от специальности подготовки экономистов учебными планами предусмотрены разные дисциплины для изучения систем и технологий: «Информационные системы в экономике», «Автоматизированные информационные системы в экономике», «Информационные технологии в управлении», «Автоматизированные информационные технологии в экономике», «Автоматизация решений учетных и аналитических задач» и др.

Информатизация общества и разработанная программа «Создание Единой системы информационного обеспечения АПК России» (ЕСИО АПК) обязывают наших выпускников быть грамотными пользователями и, возможно, принимать участие в развитии данной системы.

ЕСИО АПК осуществляет интегральную взаимоувязанную информационную поддержку всех процессов федерального, регионального и муниципального уровней государственного управления и регулирования в сфере АПК.

В учебном процессе университета по вышеназванным курсам используются разные автоматизированные информационные системы и технологии, решающие широкий спектр управленческих задач.

Для оперативного уровня управления используем системы обработки данных (СОД) – программный комплекс *1С:Предприятие 7.7*, обеспечивающий учетную функцию управления. В данном программном продукте основной технологией обработки данных является OLTP-технология (Online Transaction Processing).

Следующая группа программных продуктов – *1С:Предприятие 8.1*, *БЭСТ 5*, *БЭСТ-Маркетинг*, *БЭСТ-Финансы*. Это информационные сис-

темы управления (ИСУ), реализующие планирование, учет, анализ, контроль, применяются для функционального (тактического) уровня управления. В них наряду с OLTP-технологиями используются OLAP-технологии (Online analytical processing).

Кафедрой экономической кибернетики накоплен большой опыт преподавания автоматизированных информационных систем для оперативного и функционального уровней управления на основе вышеперечисленных и других программных продуктов.

Современные условия деятельности организаций и разных структур управления в век кибер-экономики диктуют более широкое использование систем поддержки принятия управленческих решений (СППР). На стратегическом уровне можно использовать такие СППР, как *Deductor, SAS Enterprise Guide*, работающие на основе OLAP-технологий и хранилищ данных. Хранилища данных могут заполняться из баз данных СОД и ИСУ.

Аналитическая платформа *Deductor* компании BaseGroup Labs применяется в университете с 2006 г. За истекший период с аналитической платформой познакомились более 1000 студентов разных специальностей: «Математические методы в экономике», «Бухгалтерский учет, анализ и аудит», «Финансы и кредит», «Экономика и управление на предприятии (по отраслям)», «Профессиональное обучение (экономика и управление)», а также магистры, аспиранты и преподаватели сельскохозяйственных вузов.

Студенты начинают работать с аналитической платформой *Deductor*, в зависимости от специальности, со второго по пятый курс.

Для проведения лабораторных занятий разработаны учебные пособия с реализацией интерактивного способа обучения «сквозная задача» средствами *Deductor Studio* и *SAS Enterprise Guide*. Изложенные в пособиях задания обеспечивают выполнение разнообразных функций, включающих сбор данных из различных источников, преобразование и загрузку их в хранилище, хранение информации, получение отчетности, создание произвольных запросов, многомерный анализ и др. Выбор конкретных заданий зависит от курса обучения, степени подготовки студентов и часов, предусмотренных рабочими программами.

В практической части пособия реализованы межпредметные и внутрипредметные связи. Студенты закрепляют знания по информационным и телекоммуникационным технологиям, математическим методам, статистике, анализу, прогнозированию и планированию и др.

Внутрипредметные связи обеспечивают закрепление знаний по терминологии экономической информации и соответствию ее терминологии OLAP-технологий, теории проектирования структурных элементов

АИС.

При выполнении заданий обращается внимание на научную логику построения, взаимосвязь информационных систем *1С: Предприятие*, *БЭСТ 5* и аналитических платформ *Deductor*, *SAS*.

Большое внимание уделяем организации самостоятельной работы студентов. Так, студенты специальности «Математические методы в экономике» уже на третьем курсе в рамках функционирующего на кафедре научного студенческого кружка «Современные информационные технологии и системы» закрепляют теоретические знания по использованию математических методов и различных вариантов их реализации. Кроме того, в рамках кружка студентами ведутся исследования с использованием статистических методов: кластеризации, корреляционного, регрессионного анализа, прогнозирования временных рядов, что позволяет закрепить теоретические знания и практические навыки по дисциплине «Математическая статистика».

При выполнении индивидуальных заданий на производственную практику после третьего и четвертого курсов студенты специальности «Математические методы в экономике» собирают информацию по объекту исследования в соответствии с тематикой выпускной квалификационной работы для применения аналитической платформы *Deductor*.

Результаты научных исследований студенты представляют в форме выступлений и стенд-докладов на вузовских и межвузовских научных конференциях, курсовых и выпускных квалификационных работ.

Исследования студентов с использованием аналитической платформы *Deductor* были отмечены призовыми местами на конференциях. Наши студенты принимали участие в межвузовских научно-практических конференциях, проводимых в Государственном университете – Высшая школа экономики, Московском государственном университете печати, Московском государственном университете экономики, статистики и информатики и др.

Аналитическая платформа *Deductor* используется аспирантами кафедры и факультета при подготовке диссертационных работ по специальностям 08.00.05 «Экономика и управление народным хозяйством (экономика, организация и управление предприятиями, отраслями, комплексами – АПК и сельское хозяйство)» и 08.00.13 «Математические и инструментальные методы экономики».

В перспективе планируется расширение применения аналитической платформы *Deductor* в учебном процессе и научных исследованиях.

С сентября 2010 г. на кафедре экономической кибернетики предусматривается подготовка магистров по программе «Информационное обеспечение управления АПК» направления 080500.68 «Менеджмент».

Использование аналитической платформы возможно и при реализации других магистерских программ (опыт имеется).

Важность применения систем поддержки принятия управленческих решений отмечена Департаментом научно-технологической политики и образования Министерства сельского хозяйства Российской Федерации. Кафедрой экономической кибернетики совместно с департаментом подготовлена «Программа повышения квалификации профессорско-преподавательского состава высших учебных заведений, подведомственных Минсельхозу России по теме «Системы поддержки принятия управленческих решений». Программа рассчитана на 72 часа обучения.

Таким образом, использование в учебном процессе новых информационных и телекоммуникационных технологий, включая аналитическую платформу Deductor, обеспечивает качественное преподавание дисциплин, повышает привлекательность их для студентов и тем самым реализуется задача подготовки специалистов для современного информационного общества в соответствиии с государственными образовательными стандартами второго и третьего поколений.

ОБУЧЕНИЕ СТУДЕНТОВ НАПРАВЛЕНИЯ «ПРИКЛАДНАЯ ИНФОРМАТИКА» СОВРЕМЕННЫМ ИНСТРУМЕНТАМ И ТЕХНОЛОГИЯМ АНАЛИЗА ДАННЫХ

***Прокопенко Н.Ю.**, доцент Нижегородского государственного архитектурно-строительного университета, г. Нижний Новгород*

Процесс повсеместной информатизации приводит к накоплению огромных объёмов данных во всех сферах жизни общества – от науки, бизнеса и производства до образования и здравоохранения. Принятие решений в современных условиях – процесс, требующий обработки гигабайтов информации, обусловил возрастание роли *информационно-аналитических систем*, под которыми понимают комплекс аппаратных, программных средств, информационных ресурсов, методик, которые используются для обеспечения автоматизации аналитических работ в целях обоснования принятия управленческих решений и других возможных применений.

Задачами любой информационно-аналитической системы являются

эффективное хранение, обработка и анализ данных. В связи с большим объёмом и сложностью аспект проблемы собственно анализа имеет два направления: оперативный анализ данных – On-Line Analytical Processing (OLAP), интеллектуальный анализ информации – Data Mining.

Основной задачей оперативного или OLAP-анализа является быстрое (в пределах секунд) извлечение необходимой аналитику или лицу, принимающему решения, информации, требуемой для обоснования или принятия решения.

Интеллектуальный анализ данных (ИАД) представляет собой аналитический процесс исследования человеком большого объема информации с привлечением компьютерных технологий автоматизированного поиска в исходных данных неочевидных закономерностей, что расширяет возможности аналитика, позволяя не только проверять имеющиеся гипотезы, но и генерировать новые, априорно не прогнозируемые исследователем.

Методы ИАД стали весьма широко и эффективно применяться в связи с бурным развитием в последнее десятилетие самих методик Data Mining и соответствующих инструментальных средств. Они находят применение в тех ситуациях, когда обычные методы анализа трудно или невозможно применить из-за отсутствия сведений о характере или закономерностях исследуемых процессов, взаимозависимостях явлений, фактов, поведении объектов и систем из различных предметных областей, в том числе в социальной и экономической.

Задачами интеллектуального анализа являются:

- выявление взаимозависимостей, причинно-следственных связей, ассоциаций и аналогий, определение значения фактора времени, локализация событий или явлений по месту.
- классификация событий и ситуаций, материальных и других объектов по совокупностям признаков, определение профилей различных факторов;
- прогнозирование событий, хода процессов;
- оценка эффективности деятельности, проектов.

Если традиционный анализ данных опирался, в первую очередь, на методы прикладной статистики, то новое направление обработки данных – ИАД – использует технологии нейронных сетей, генетических алгоритмов, нечеткой логики и другие инструменты современной математики, логики, теории искусственного интеллекта. Методы Data Mining призваны не только описать зависимости и взаимосвязи, но и объяснить их. Они не налагают априорных гипотез на данные, не навязывают реальности заранее выбранных шаблонов, работая даже тогда, когда данные (а значит и описываемая ими предметная область) имеют сложную, запу-

танную структуру.

В настоящее время большую актуальность приобретает профессиональное владение современными технологиями анализа данных. Программы нечеткой логики, генетических алгоритмов, нейронных сетей предназначены не для программистов, а для управленцев, менеджеров, инженеров. Они являются незаменимыми помощниками при принятии решений, анализе ситуации, прогнозировании развития событий. Именно поэтому актуальной является задача подготовки нового поколения специалистов, способного на практике осознано применять новые информационные технологии, составной частью которых являются интеллектуальные средства обработки информации.

Для того чтобы студенты направления «Прикладная информатика» могли свободно пользоваться современными IT-технологиями, мы используем три пути.

1. *Чтение специальных курсов*, посвященных ИАД: «Интеллектуальные информационные системы», «Компьютерные технологии в науке, образовании и производстве». Главная цель этих дисциплин – получение базовых знаний в области интеллектуальных информационных систем, изучение методов и средств интеллектуального анализа данных, приобретение навыков работы со средствами Data Mining.

2. *Внедрение в другие курсы*: «Базы данных», «Эконометрика», «Логистика», «Методы прогнозирования». Активное использование межпредметных связей перечисленных учебных дисциплин играет важную роль в повышении практической и научно-теоретической подготовки студентов, существенной особенностью которой является овладение ими обобщенным характером познавательной деятельности.

3. *Индивидуальная работа со студентами* (НИР, курсовые и дипломные работы).

Для научно-исследовательской работы студентов, как для одной из форм учебного процесса, характерно удачное сочетание обучения и практики. В рамках этой работы студент приобретает сначала основные навыки исследовательской работы, а затем начинает воплощать теоретические знания в научных исследованиях, выполняемых на кафедре. Ее основные формы – участие в факультетских, университетских, межвузовских научных конференциях и студенческие публикации, а также пилотные проекты и исследовательская работа в области учета и анализа данных.

Подчеркнем, что интеллектуальный анализ данных невозможно реализовать без специализированных пакетов программ, реализующих эвристические алгоритмы выявления закономерностей, релевантных как накопленным данным, так и целям их обработки. ИАД предполагает на-

личие *данных* (числовых, текстовых или других), цели, определяющей вид искомой закономерности (ассоциации, классификации, кластеризации или др.), *математического аппарата*, способного решить задачу поиска определенного вида закономерности, и *программного инструмента*, реализующего соответствующий математический метод.

Для обучения студентов в качестве универсальной моделирующей среды для создания прикладных решений в области анализа данных нами была выбрана аналитическая платформа *Deductor* (отечественная разработка компании BaseGroup Labs). Выбор этого программного продукта был обусловлен рядом причин:

- наличие учебной версии;
- наличие в *Deductor Academic* самых современных методов извлечения, манипулирования, визуализации данных, кластеризации, прогнозирования и многих других технологий интеллектуального анализа данных;
- различные методические материалы (демо-сценарии, примеры, аналогии), объясняющие сложные понятия, термины, принципы работы тех или иных методов анализа и механизмов построения аналитических решений;
- множество способов визуализации данных (кроме традиционных таблиц, графиков и диаграмм используются специальные визуализаторы такие как граф нейронной сети, сети Кохонена, деревья решений, рок-диаграмма, таблица сопряженности);
- доступность для освоения пользователями, имеющими разные уровни компьютерной и математической подготовки.

Конечно, знание таких дисциплин как общий курс высшей математики, теории вероятностей и математической статистики является необходимым фундаментом изучения методов Data Mining. Но в тоже время использование самообучающихся методов и мастеров для настройки, множества способов визуализации данных делает современные технологии доступными широкому кругу пользователей, позволяя снизить требования к уровню математической подготовки студентов.

Важной особенностью интеллектуального анализа данных является то, что он позволяет более полно использовать способности человека, освобождая его не только от рутинных вычислений, но даже от формулировки гипотез (естественно, при наличии «сильной» интеллектуальной системы, оснащенной «хорошим» математическим аппаратом, позволяющим реализовать методологию генерации и отбора наиболее интересных гипотез). Однако ИАД не решает задачи за аналитика, а всего лишь служит инструментом, который способствует поиску нетривиальных решений содержательных задач. Кроме того, ИАД предполагает со-

вместное использование различных методов и алгоритмов в процессе анализа эмпирических данных. В данном контексте особое значение приобретают математические знания, навыки компьютерной реализации различных методов анализа данных и корректной интерпретации полученных результатов.

Условия успеха в интеллектуальном анализе данных:

- ясность в представлении цели анализа;
- подготовка существенных для проводимого исследования данных;
- правильный выбор методов и программных средств;
- квалифицированное и тщательное выполнение методов анализа;
- решение о применении результатов анализа.

Свободно распространяемая академическая версия аналитической платформы *Deductor* имеет большое значение для подготовки специалистов в области IT-технологий, в сфере экономики, маркетинга и менеджмента, так как очевидно, что теоретических знаний молодым специалистам недостаточно, нужно уметь применять полученные знания к решению прикладных задач. Добиться этого можно, только проведя достаточно много времени за компьютером, осваивая соответствующие программные продукты. Применение *АП Deductor* позволяет двигаться постепенно от простых задач к более сложным, получая отдачу на каждом шаге. Любой набор данных на каждом этапе анализа можно визуализировать каким-либо доступным способом или несколькими способами, поскольку визуализация помогает интерпретировать построенные модели. Такой подход значительно облегчает восприятие студентами содержания решаемых бизнес-задач. Благодаря доступности и наглядности методов анализа данных, реализованных в *Deductor*, студентам остается главным образом творческая работа: изучение предметной области, выбор методов решения, интерпретация результатов.

Возможность применять в конкретных ситуациях, при решении реальных задач знания и умения, полученные во время учебных теоретических и практических занятий, появляется у студентов с момента прохождения производственной и преддипломной практики, целью которой является освоение самых современных информационных технологий, ознакомление и исследование новейших тенденций и разработок в области информационных технологий по созданию систем поддержки принятия решений.

Цель производственной практики – изучение опыта создания и применения конкретных информационных технологий и систем информационного обеспечения для решения реальных задач организационной,

управленческой и научной деятельности в условиях конкретных производств, организаций или фирм.

Цель преддипломной практики – сбор необходимого материала для выполнения выпускной квалификационной работы. Осуществляется углубление и закрепление полученных студентами теоретических и практических навыков в различных областях деятельности информатика-экономиста (организационно-управленческая, проектно-технологическая, маркетинговая, экспериментально-исследовательская, консалтинговая, аналитическая и эксплуатационная); приобретение ими навыков практического решения информационных задач на конкретном рабочем месте.

Значительное внимание при прохождении практики уделяется таким вопросам, как:

- учет и автоматизацию каких бизнес-процессов необходимо наладить в первую очередь;
- в каких случаях и для решения каких задач применять тот или иной инструментарий;
- как подготовить данные для анализа;
- как интерпретировать результаты для выработки эффективных управленческих решений;
- как комбинировать методы анализа;
- как быстрее получить нужный результат, не снижая качества.

Проблемой при решении этих задач является необходимость популяризации новых информационных технологий, так как сотрудники во многих компаниях зачастую задавлены текучкой, и не успевают следить за новостями на IT-рынке. Достаточно часто студенты-практиканты сталкиваются еще и с тем фактом, что предприятия при выборе информационной системы ставят во главу угла не ее функциональные возможности, а исключительно стоимость. Они привыкли работать, используя уже имеющиеся программные продукты, и не хотят внедрять новые (делают простейшую аналитику в *Excel* и считают, что этого им достаточно). Кроме того, поскольку технологии ИАД являются мультидисциплинарной областью, для разработки приложений, включающих методы Data Mining, необходимо задействовать специалистов из разных областей, а также обеспечить их качественное взаимодействие. Однако специалистов по Data Mining, которые бы хорошо разбирались в бизнесе, еще очень мало. К сожалению, на предприятиях, где студенты проходят практику, не всегда есть возможность получить консультацию у экспертов – руководителей и специалистов организации, которые, ссылаясь на производственную загруженность, в лучшем случае предоставляют толь-

ко данные и ждут готовых результатов.

Для того чтобы готовить интересные, имеющие реальное практическое значение выпускные квалификационные работы, необходимо взаимодействие вуза с организациями, использующими или внедряющими современные интеллектуальные информационные системы. У нас уже есть положительный опыт – это сотрудничество с компаниями BaseGroup Labs и ее партнера BIGroup Labs (г. Нижний Новгород, www.bi-grouplabs.ru), основной образовательной целью которых является подготовка грамотных специалистов, понимающих потребности бизнеса и умеющих применить современные информационные технологии для их удовлетворения. Огромные усилия компания BaseGroup Labs вкладывает в образовательную программу: открыт образовательный портал и запущена полноценная система дистанционного обучения, где предложены самые современные курсы обучения и сертификации для преподавателей вузов по анализу данных; выпустила академическую версию аналитической платформы *Deductor Academic*, на основе которой изучаются все практические аспекты технологий анализа данных.

Используя функциональность и аналитические возможности платформы *Deductor*, студенты в ННГАСУ готовят курсовые и дипломные работы, которые отличаются новизной подходов и практической значимостью. Так, например, ими были получены интересные аналитические решения в области бизнеса и управления, оформленные в виде выпускных квалификационных работ: «Разработка интегрированной информационно-аналитической системы поддержки принятия решений регионального управления» (Д.В. Власенко), «Разработка интегрированной системы информационно-аналитического обеспечения деятельности Нижегородской областной детской клинической больницы» (О.В. Рабынина), «Применение современных информационных технологий и интеллектуальных методов анализа в задаче оценки недвижимости» (Т.В. Медведева), «Разработка автоматизированной информационной системы принятия инвестиционных решений на базе АП *Deductor*» (И.В. Ильин).

Целью первых двух работ является разработка методологии использования различных информационных систем (офисных, интеллектуальных, геоинформационных) и новых информационных технологий для интегрированной обработки экономической и медицинской информации в системах поддержки принятия решений. В дипломных проектах рассматривались теоретические аспекты и практическая реализация интегрированной информационно-аналитической системы поддержки принятия решений, которая состоит из 4 подсистем.

1. Подсистема сбора и хранения данных реализована средствами *MS Access* и *АП Deductor*. Создание интегрированного хранилища дан-

ных, а также организация обработки накопленной в *MS Access* информации было реализовано на базе *Deductor Warehouse*.

2. Подсистема очистки и подготовки данных к анализу и построению аналитических моделей реализована в *Deductor Studio*, где имеются все необходимые обработчики для проведения аудита данных, подготовки данных к анализу и построению аналитических моделей. Аудит данных включает в себя: проверку и устранение дубликатов и противоречий, обработку пропусков, выявление выбросов и фильтрацию.

3. Подсистема моделирования и прогнозирования реализована на базе *Deductor Studio*. Она включает несколько прогностических и классификационных моделей. С помощью обработчиков в *Deductor* был проведен корреляционный анализ отобранных показателей и построены модели прогноза: линейная регрессионная модель и нейросетевой прогноз временного ряда.

4. Подсистема аналитической отчетности и географического отображения данных реализована средствами *АП Deductor* и геоинформационной системы *MapInfo*. Подсистема построения отчетов предполагает графическое представление данных в виде графиков, диаграмм, карт Кохонена, а также предполагает нанесение результатов анализа и прогнозирования рассматриваемых показателей на карту Нижегородской области для сравнительного анализа состояний муниципальных районов.

Аналитическая отчетность в *Deductor* обеспечивает быстрый доступ к результатам анализа, не требуя от пользователя навыков анализа данных и работы в АП. При работе с отчетами пользователь не видит сценарий анализа данных, ему доступны только конечные результаты (выдержки) из работы аналитика. Отчеты построены в виде древовидного иерархического списка, каждым узлом которого является отдельный отчет или папка, содержащая несколько отчетов. Каждый узел дерева отчетности связан со своим узлом в дереве сценария. Для каждого отчета были построены OLAP-кубы, а также настроены другие способы отображения: гистограммы, кросс-таблицы, кросс-диаграммы.

В дипломном проекте *«Применение современных информационных технологий и интеллектуальных методов анализа в задаче оценки недвижимости»* описываются новые подходы к решению задач оценки недвижимости. При помощи аналитической платформы *Deductor* был построен сценарий, решающий задачи классификации и прогнозирования стоимости жилья. Задачи классификации решены при помощи моделей дерева решений и нейронной сети, задачи прогнозирования – при помощи модели нейронной сети и модели множественной регрессии. Также было проведено сравнение модели на основе показателей оценки их качества.

Основным результатом дипломного проекта *«Разработка автоматизированной информационной системы принятия инвестиционных решений на базе АП Deductor»* является вывод о том, что данная аналитическая платформа может эффективно справляться с задачами отчистки и обработки информации, а благодаря встроенному полнофункциональному нейросетевому аппарату, может быть использована для решения широкого круга задач оценки и прогнозирования различных финансовых величин.

Опыт преподавания методов Data Mining показывает, что большинство студентов, ориентированных после окончания университета работать по специальности, интересуются возможностями технологий ИАД и согласны прилагать усилия, чтобы разобраться в их тонкостях, несмотря на сложность теоретических и прикладных аспектов интеллектуальных вычислений. Мотивацией к изучению ИАД, как нам представляется, может служить ознакомление с результатами его применения. Наша собственная практика применения ИАД в дипломных проектах дает основания с уверенностью утверждать, что ИАД является очень полезным, инструментом познания действительности, решения разнообразных бизнес-задач. Освоение методов ИАД будущими специалистами по информационным технологиям является важным шагом в повышении уровня их профессиональной культуры.

ПРОБЛЕМЫ ОБУЧЕНИЯ СТУДЕНТОВ КОНЦЕПТУАЛЬНОМУ АНАЛИЗУ ДАННЫХ

***Шамсутдинова Т.М.**, доцент Башкирского государственного аграрного университета, г. Уфа*

Концептуальный анализ данных является одним из основных этапов проектирования интеллектуальных информационных систем, основанных на концепции использования баз знаний. К интеллектуальным системам такого типа относятся системы с интеллектуальным интерфейсом, экспертные системы, самообучающиеся и адаптивные информационные системы [1].

Как известно, база знаний – это хранилище единиц знаний, описывающих атрибуты и действия, связанные с объектами проблемной области. Для представления единиц знаний в них обычно используются правила, описывающие поведение и взаимодействие исследуемых объектов.

Цель концептуального анализа данных при построении баз знаний – провести содержательный анализ проблемной области, выявить в ней основные понятия и их взаимосвязи.

Можно выделить следующие основные уровни концептуального анализа данных [2, 3].

Объектно-структурный уровень. На данном уровне концептуального анализа данных разрабатывается *структурная* модель исследуемой предметной области. Данная модель описывает структуру предметной области как совокупности взаимосвязанных объектов, отражает фактуальное знание о составе объектов, их свойствах и связях. Элементарной единицей структурного знания является факт, описывающий одно свойство или одну связь объекта и представляемый в виде: предикат (Объект, Значение). Для представления данных моделей используются средства ER-описаний, например, модель «Сущность-Связь».

Функциональный уровень. *Функциональная* модель предметной области отражает основные функциональные связи между объектами, описывает преобразования фактов и зависимости между ними, показывает, как определенные факты образуются из других. В качестве единицы функционального знания используется зависимость фактов в виде: A → B. Формами представления функциональных моделей являются IDEF-диаграммы, деревья целей, графы И-ИЛИ.

Поведенческий уровень. *Поведенческая* модель предметной области рассматривает взаимодействия объектов во временном аспекте. Данная модель отражает изменение состояний объектов в результате возникновения некоторых событий, влекущих за собой выполнение определенных действий. Для представления поведенческих моделей используются описания потоков событий, например, DFD-диаграммы.

К прикладным методикам концептуального анализа предметной области при этом можно отнести:

- выявление корреляционных и регрессионных зависимостей между данными;
- анализ данных методом главных компонент;
- кластерный анализ (в том числе с применением самоорганизующихся карт Кохонена);
- построение деревьев решений;
- анализ данных с применением нейронных сетей;
- построение ассоциативных правил и др.

Для обучения студентов методикам концептуального анализа данных в Башкирском государственном аграрном университете используется аналитическая платформа *Deductor*, позволяющая реализовать основ-

ные стратегии извлечения знаний. Опыт использования данного программного продукта на учебных занятиях составляет три года, занятия ведутся на старших курсах специальности 080801 «Прикладная информатика (в экономике)» в рамках дисциплины СД.Ф.02 «Интеллектуальные информационные системы».

В ходе занятий студенты выполняют цикл лабораторных работ, связанных с изучением методики интеллектуального анализа данных на основе построения деревьев решений, нейронных сетей, карт Кохонена и т.д. Далее студенты выполняют курсовую работу, связанную с проектированием экспертных систем, в которой они должны провести концептуальный анализ данных с целью построения базы знаний своей экспертной системы.

К основным этапам проектирования экспертных систем традиционно относят такие этапы как:

1. идентификация проблемной области;
2. концептуализация проблемной области;
3. формализация базы знаний;
4. реализация экспертной системы;
5. тестирование;
6. внедрение и опытная эксплуатации.

Из всех перечисленных этапов наибольшее затруднение у студентов, как правило, вызывает именно этап концептуализации проблемной области. На данном этапе студентами совершается достаточно большое количество ошибок, существенно снижающих ценность выполненной работы.

Исходя из опыта преподавательской работы в вузе, была сформулирована следующая система оценки качества концептуального анализа данных в курсовых работах студентов специальности «Прикладная информатика (в экономике)» по дисциплине «Интеллектуальные информационные системы». Для оценки объема и качества проведенного концептуального анализа предлагается использовать следующую систему критериев:

- актуальность темы работы;
- четкость формулирования цели концептуального анализа данных;
- полнота составления исходной выборки данных, включая количество рассматриваемых факторов;
- уровень структурированности рассматриваемых факторов;
- наличие системного подхода к рассмотрению факторов;
- четкость формулирования гипотез;

- количество использованных методик концептуального анализа;
- качество построенных моделей концептуального анализа данных;
- степень согласованности построенных моделей анализа данных;
- степень глубины проведенного концептуального анализа;
- полнота рассмотрения отдельных аспектов предметной области;
- степень соответствия между результатами концептуального анализа и строящейся далее базой знаний;
- шаблонность работы и объем «заимствований» по системе «антиплагиат»;
- выполнение графика сдачи этапов работы.

К наиболее типичным ошибкам студентов при проведении концептуального анализа данных можно отнести бессистемность и неполноту исходной выборки данных, несогласованность построенных моделей анализа, низкую степень глубины проведенного концептуального анализа, слабое соответствие между результатами концептуального анализа данных и строящейся далее базой знаний. Достаточно часто проведенный анализ носит поверхностный характер и плохо раскрывает связи между объектами проблемной области.

Довольно много проблем возникает у студентов при проведении кластерного анализа данных. Не всем студентам удается выделить общие для кластеров свойства объектов, четко сформулировать критерии отнесения объектов к тому или иному кластеру. Бывают случаи, когда выбор входных и выходных полей карт Кохонена, например, вообще не соответствует общей концепции анализа проблемной области.

Ряд сложностей возникает у студентов и при анализе данных с использованием нейронных сетей. Причем здесь есть как проблемы практического характера (непонимание принципов нормализации полей, выбора топологии сети, настройки параметров ее обучения), так и проблемы аналитического характера, связанные с отсутствием четкого понимания итоговой цели применения нейронной сети для концептуального анализа своей проблемной области.

После этапа концептуализации проблемной области должен реализовываться этап формализации базы знаний, на котором формализуются логические правила базы знаний экспертной системы. Многие студенты, нарушая целостность процесса проектирования своей системы, составляют правила, не согласующиеся с результатами проведенного концептуального анализа данных. При этом вводятся совершенно новые факторы, не рассмотренные ранее и не имеющие своего места в уже сформированной структуре поля знаний.

В заключение можно только добавить, что проблема обучения студентов навыкам концептуального анализа данных носит комплексный характер. Для ее решения необходимо повышать уровень абстрактности мышления студентов, формировать у них творческий подход к решению исследовательских задач, развивать аналитические способности и навыки системного подхода к анализу данных.

Литература

1. Андрейчиков А.В. Интеллектуальные информационные системы. – М.: Финансы и статистика, 2003.
2. Тельнов Ю.Ф. Интеллектуальные информационные системы. – М: Московский международный институт эконометрики, информатики, финансов и права, 2004.
3. Гаврилова Т.А., Хорошевский В.Ф. Базы знаний интеллектуальных систем. – СПб: Питер, 2000. – 384 с.

АНАЛИТИЧЕСКАЯ ПЛАТФОРМА DEDUCTOR В МОДЕЛИРОВАНИИ ПРИНЯТИЯ ОПЕРАТИВНЫХ УПРАВЛЕНЧЕСКИХ РЕШЕНИЙ

***Павленко Л.А.**, доцент Харьковского национального экономического университета, г. Харьков*
***Тарасов А.В.**, доцент Харьковского национального экономического университета, г. Харьков*

Оперативный анализ информации о деятельности предприятий любой формы собственности, вида и масштаба деятельности является основой принятия, как оперативных, так и стратегических управленческих решений.

Обучение вопросам анализа данных студентов направления «Компьютерные науки» на кафедре информационных систем ХНЭУ ориентировано на ведущую роль эксперта, как аналитика или лица принимающего решение (ЛПР) в диалоге с системой поддержки принятия решений (СППР). СППР рассматривается как интерактивная прикладная система, которая обеспечивает конечного пользователя, быстрым, удобным доступом к данным и моделям с целью принятия решений в сложных ситуациях [1]. При этом конечных пользователей СППР относят к одной из трех категорий: низшее, среднее и высшее звено руководящих работни-

ков. Первые используют, как детализированные, так и слабо агрегированные данные. Вторые используют слабо и высоко агрегированные данные, относящиеся к сфере деятельности отдельного подразделения, для принятия тактических решений. Третьи использует высоко агрегированные данные, отображающие деятельность организации в целом, для принятия стратегических решений [2].

Моделирование процедур принятия решений неразрывно связано с концепцией OLAP – технологии обработки данных, сформулированной Е. Коддом и связанной с ней проблемой организации баз данных [3]. Изучение вопросов проектирования баз данных (БД) на кафедре ориентировано на принцип разделения базы на транзакционную и аналитическую компоненты.

Транзакционной часть БД является первым шагом в создании хранилища данных и используется для разработки разнообразных информационных модулей учета выполнения тех или иных бизнес-процессов. Характерными чертами транзакционной базы являются:

1. реляционная структура (преимущественно);
2. накопление значительных объемов фактических данных;
3. выполнение операций добавления, удаления, редактирования записей;
4. отсутствие агрегированных (вычисленных) данных.

Аналитическая часть БД используется при выполнении оперативного анализа информации и разработке моделей для систем поддержки принятия решений. Характерными чертами аналитической БД являются:

1. многомерная структура (поддержка моделей: MOLAP, ROLAP, HOLAP);
2. хранение агрегированных данных;
3. отсутствие операций удаления и редактирования агрегатов данных и накопление их, как правило, по хронологии.

Сегодня для моделирования бизнес-процессов и разработки схем данных существует множество CASE-инструментов. На этапе, предшествующем разработке схемы данных, связанном с описанием бизнес-процессов, используются пакеты: структурного моделирования BPwin, объектно-ориентированного моделирования Rational Rose, ARIS и другие.

При разработке логической и физической схем данных используются такие пакеты, как *ERwin*, *Oracle Designer*, *Visio* и множество других. Инструментальная среда разработки многомерных моделей является неотъемлемой частью любой мощной СУБД. Но далеко не все фирмы могут позволить себе приобретение и поддержку легальной версии такого инструмента как *Oracle Designer*. Особую актуальность для фирм ма-

лого и среднего бизнеса имеет вопрос использования недорого профессионального инструмента, обладающего средствами обмена данными и результатами моделирования с широко распространенными инструментами *MS Office*.

Кроме того, изучение студентами такого инструментария позволяет приобрести навыки и достичь необходимого уровня компетенции в вопросах манипулирования данными и многомерного анализа. Таким простым и удобным в использовании инструментом является аналитическая платформа *Deductor* [4].

В данной работе приведено сравнение возможностей различных OLAP-схем и инструментов организации хранения и анализа данных на примере исследования популярности у покупателей продукции различных фирм-производителей. Результаты анализа необходимы руководству торгующей организации для принятия решений о дальнейшем сотрудничестве с фирмами-производителями.

Подобные задачи решаются в ходе проведения лабораторных работ со студентами. В данном случае решение выполнено средствами *ERwin*, *MS Access*, *MS Excel* и *Deductor Academic*.

Даны результаты фиксации сумм платежей, выручаемых торгующей организацией от ежедневных продаж продукции трех фирм-производителей: «АРГО», «НОВЬ», «БЛЕСК». Данные накапливались в течение нескольких лет в ROLAP-модели «Снежинка». Учет выполняется по разным производителям, годам, месяцам, дням недели, конкретным датам. Причем продажи осуществляются ежедневно, без выходных.

Схема данных ROLAP, разработанная в среде *ERwin*, приведена на рис. 1. Та же схема, полученная инструментом прямого инжиниринга физической модели данных пакета *ERwin* в среду *MS Access*, приведена на рис. 2.

В модели различаются таблицы измерений: *Год*, *Месяц*, *День недели*, *Дата*, *Фирма* и таблица фактов: *Платеж*.

Достоинствами ROLAP-модели являются следующие.

1. Возможность реализации в среде реляционной СУБД и выполнения SQL запросов с последующим построением диаграмм и графиков анализа данных.
2. Модель позволяет выполнить операции «Подъема» и «Спуска» по измерениям и получить «Срез» данных по произвольно выбранным измерениям.

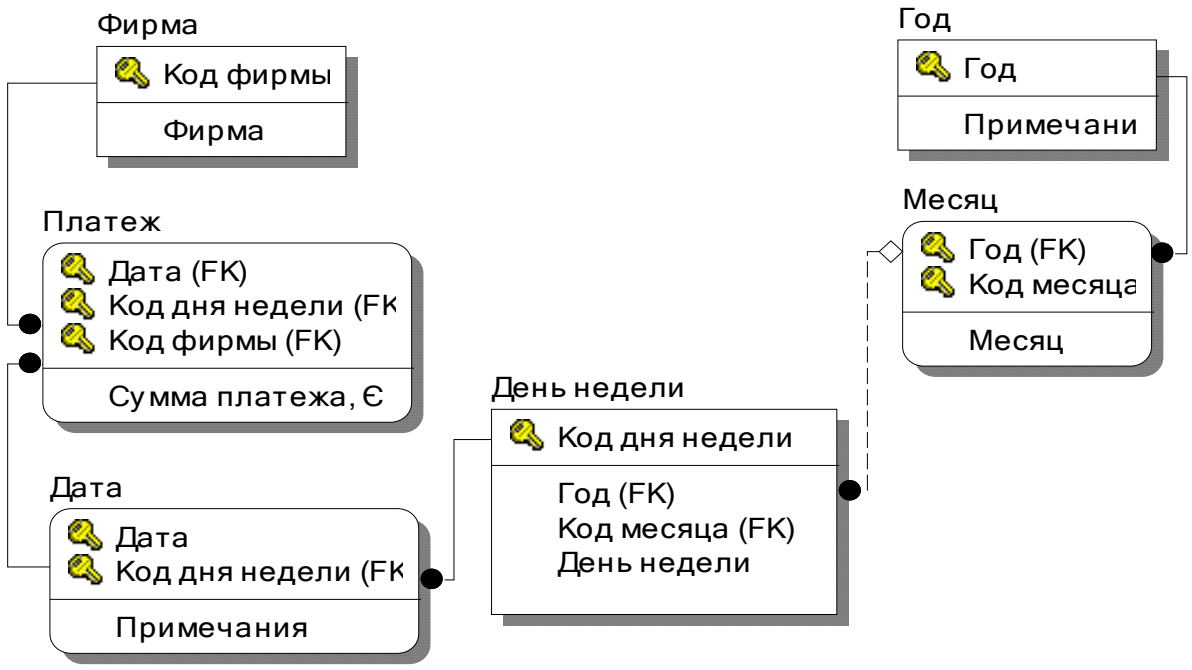


Рис. 1. Схема данных ROLAP в среде ERwin (нотация IDEF1X)

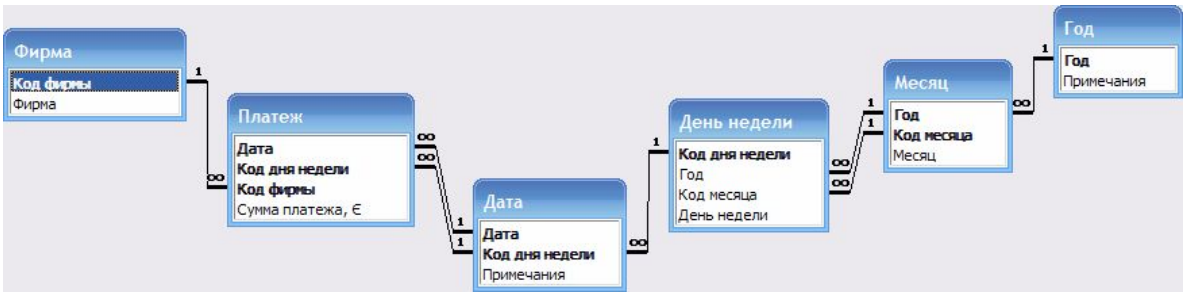


Рис. 2. Схема данных ROLAP в среде Access

Например, зафиксировав определенный год и выполнив операцию естественного соединения таблиц *Фирма*, *Платеж*, *Дата*, *День недели*, *Месяц*, можно получить информацию о суммах, вырученных от реализации продукции по датам и фирмам за все дни недели и месяцы этого года.

Недостатком модели является (особенно это касается календаря дат) избыточность данных. Например, в таблице *Месяц* дублируются наименования месяцев, в таблице *День недели* дублируются наименования дней недели. Цепочка таблиц *Год*, *Месяц*, *День недели*, *Дата* представляют собой «древовидный» громоздкий календарь, неудобный, как при разработке структуры, так и в процессе работы с ним. В отличие от него календарь, ориентированный на реляционную модель, состоит из нормализованных таблиц и позволяет выполнять все операции поиска и фиксации необходимых дат.

Альтернативой ROLAP является кубическая модель. Простым, доступным и удобным средством построение такой модели является инструмент *MS Excel* «Сводная таблица».

Результат соединения таблиц схемы рис. 2 был экспортирован в среду Excel. На рис. 3 представлен фрагмент полученной по этим данным сводной таблицы. Модель позволяет фиксировать: год, месяц, дату, день недели, фирму, как в отдельности, так и поочередно, по желанию пользователя, с целью анализа результатов фильтрации данных и принятия управленческих решений.

На рис. 4 приведена соответствующая этой таблице диаграмма для января 2010 года, которая перестраивается при изменении фиксации измерений. Можно поочередно фиксировать: год, месяц, дату, день недели, фирму и анализировать результаты выбора данных.

	A	B	C	D	E	F	G
1	Год	2010					
2							
3	Сумма по по			Фирма			
4	Месяц	Дата	День недели	АРГО	БЛЕСК	НОВЬ	Общий итог
5	Январь	1	Пятница	89,19	22,87	108,34	220,4
6		2	Суббота	34	45,8	96,9	176,7
7		3	Воскресенье	52	23,8	82,9	158,7
8		4	Понедельник	45,6	42,8	92,7	181,1
9		5	Вторник	67,9	36,8	92,6	197,3
10		6	Среда	61,6	39,5	93,6	194,7
11		7	Четверг	70	39	83	192
12		8	Пятница	58,3	56	82	196,3
13		9	Суббота	72	63	98	233
14		10	Воскресенье	56	53	94,6	203,6
15		11	Понедельник	69	39,5	108,9	217,4
16		12	Вторник	85	56	93	234
17		13	Среда	58,3	51	81,9	191,2
18		14	Четверг	58,3	76	85	219,3
19		15	Пятница	72	91	97,1	260,1
20		16	Суббота	73	64	93,6	230,6
21		17	Воскресенье	75	86	95	256
22		18	Понедельник	98,3	79	98,89	276,19
23		19	Вторник	78	51	93,6	222,6
24		20	Среда	82	65,6	93	240,6
25		21	Четверг	67,9	56,7	95,5	220,1
26		22	Пятница	44,78	48,1	98,9	191,78
27		23	Суббота	75	38	97	210
28		24	Воскресенье	76	64	93,6	233,6
29		25	Понедельник	59,99	52	108,3	220,29
30		26	Вторник	63	39,5	95	197,5
31		27	Среда	75	57	106	238
32		28	Четверг	78,3	39,5	98	215,8
33		29	Пятница	72	58	110	240
34		30	Суббота	86	79	120	285
35		31	Воскресенье	75	55,6	93,6	224,2
36	Январь Итог			2128,46	1669,07	2980,53	6778,06
37	Общий итог			2128,46	1669,07	2980,53	6778,06

Рис. 3. Фрагмент сводной таблицы в среде Excel

Недостатком инструмента *Excel* «Сводная таблица» является отсутствие возможности быстрого транспонирования матрицы данных, как на уровне таблицы, так и на уровне диаграммы, то есть выполнения операции «Вращение куба», что замедляет процесс анализа данных.

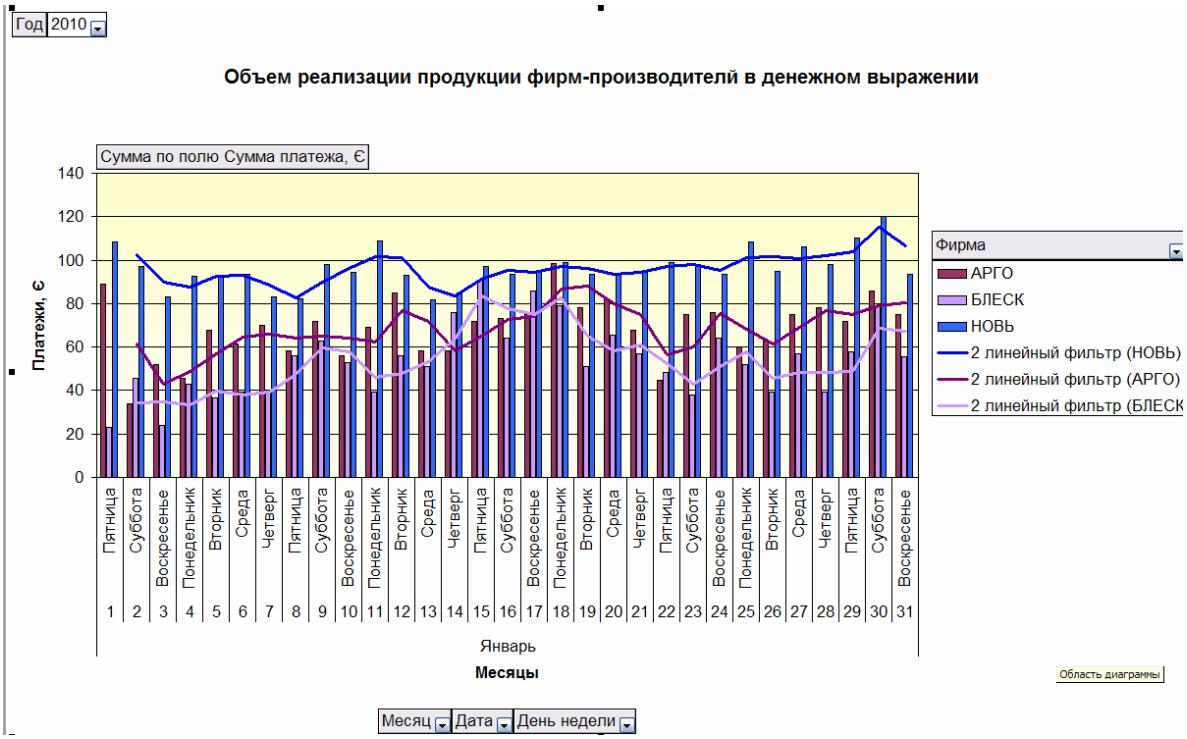


Рис. 4. Диаграмма, построенная по сводной таблице данных (рис. 3)

Для построения многомерной модели «Куб» средствами *Deductor Academic* данные таблицы рис. 2 были сохранены как текстовый файл и импортированы в среду пакета.

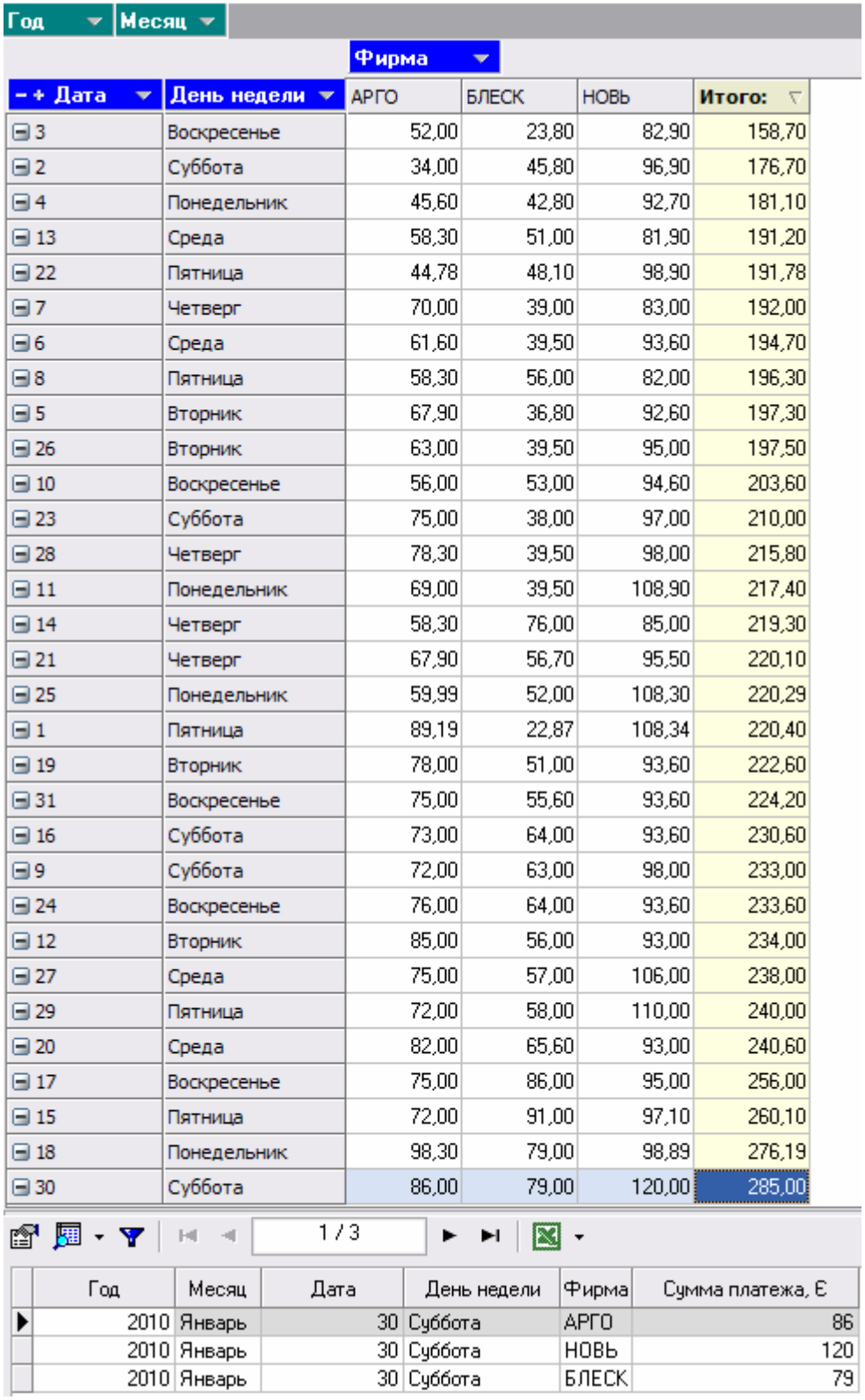
На рис. 5 приведен результат построения куба в среде *Deductor Studio Academic*.

OLAP-инструментарий аналитической платформы *Deductor* прост, понятен и удобен, как в процессе построения модели, так и в процессе анализа данных.

1. По желанию пользователя можно анализировать данные в самом кубе и во вспомогательной таблице детализации, отражающей данные определенных ячеек куба.
2. Структуру куба можно перестраивать в процессе анализа.
3. Диаграмма для данных куба оперативно вызывается на одном листе с данными (рис. 6).
4. Deductor позволяет выполнять все необходимые операции анализа данных: «Подъема», «Спуска», «Вращения» (транспонирование матрицы данных), «Среза». Причем транспонировать

матрицу можно как в поле сводной таблицы, так и поле диаграммы, что повышает оперативность анализа.

5. Возможно выполнить экспорт результатов моделирования в Word, Excel, сохранить в формате HTML.



Год | Месяц | Фирма

Дата	День недели	АРГО	БЛЕСК	НОВЬ	Итого:
3	Воскресенье	52,00	23,80	82,90	158,70
2	Суббота	34,00	45,80	96,90	176,70
4	Понедельник	45,60	42,80	92,70	181,10
13	Среда	58,30	51,00	81,90	191,20
22	Пятница	44,78	48,10	98,90	191,78
7	Четверг	70,00	39,00	83,00	192,00
6	Среда	61,60	39,50	93,60	194,70
8	Пятница	58,30	56,00	82,00	196,30
5	Вторник	67,90	36,80	92,60	197,30
26	Вторник	63,00	39,50	95,00	197,50
10	Воскресенье	56,00	53,00	94,60	203,60
23	Суббота	75,00	38,00	97,00	210,00
28	Четверг	78,30	39,50	98,00	215,80
11	Понедельник	69,00	39,50	108,90	217,40
14	Четверг	58,30	76,00	85,00	219,30
21	Четверг	67,90	56,70	95,50	220,10
25	Понедельник	59,99	52,00	108,30	220,29
1	Пятница	89,19	22,87	108,34	220,40
19	Вторник	78,00	51,00	93,60	222,60
31	Воскресенье	75,00	55,60	93,60	224,20
16	Суббота	73,00	64,00	93,60	230,60
9	Суббота	72,00	63,00	98,00	233,00
24	Воскресенье	76,00	64,00	93,60	233,60
12	Вторник	85,00	56,00	93,00	234,00
27	Среда	75,00	57,00	106,00	238,00
29	Пятница	72,00	58,00	110,00	240,00
20	Среда	82,00	65,60	93,00	240,60
17	Воскресенье	75,00	86,00	95,00	256,00
15	Пятница	72,00	91,00	97,10	260,10
18	Понедельник	98,30	79,00	98,89	276,19
30	Суббота	86,00	79,00	120,00	285,00

1 / 3

Год	Месяц	Дата	День недели	Фирма	Сумма платежа, Є
2010	Январь	30	Суббота	АРГО	86
2010	Январь	30	Суббота	НОВЬ	120
2010	Январь	30	Суббота	БЛЕСК	79

Рис. 5. Результат построения куба в аналитической платформе Deductor

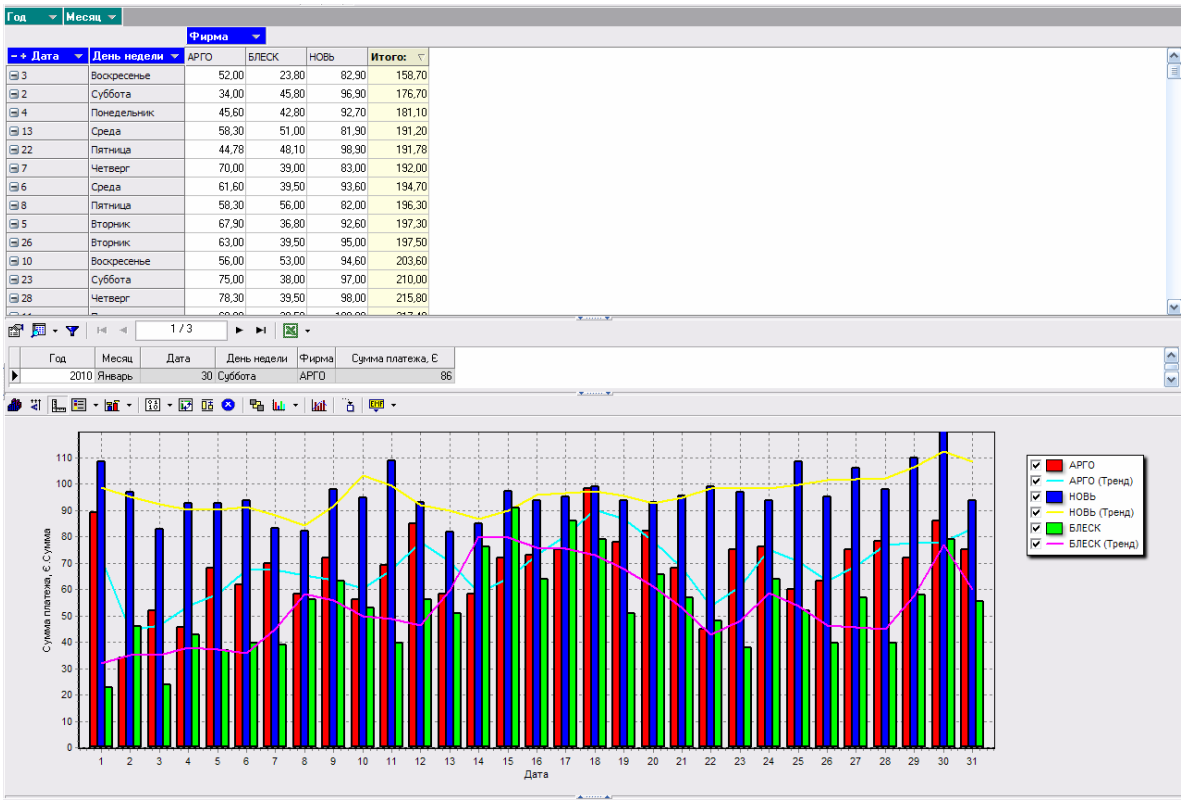


Рис. 6. Диаграмма анализа данных

В результате выполнения вращения куба и фильтрации данных на диаграмме (рис. 7) отражены платежи за приобретенную продукцию в воскресные и субботние дни января 2010 года, сгруппированные по фирмам.

Рис. 7. Диаграмма анализа отфильтрованных данных в транспонированной матрице

Полученная модель позволяет аналитику торгующей организации сделать вывод о том, что продукция фирмы «НОВЬ» пользуется большим спросом у покупателей и приносит более стабильный доход, чем продукция фирм «АРГО» и «БЛЕСК». И в выходные дни продукция фирмы «НОВЬ» пользуется большим спросом. Модель позволяет анализировать более удачные дни недели, месяцы и годы реализации продукции разных фирм и делать соответствующие организационные выводы.

В результате проведения лабораторных работ студенты приобретают навыки и умения в выполнении операций анализа и принятия управленческих решений на основе многомерной организации данных.

Несомненные достоинства средств построения OLAP-моделей аналитической платформы *Deductor* позволяют эффективно использовать ее как на занятиях со студентами, так и в качестве рабочего инструмента аналитиков фирм различных видов и уровней деятельности.

Литература

1. Андрейчиков А.В., Андрейчикова О.Н. Анализ, синтез, планирование решений в экономике. — М.: Финансы и статистика, 2002. – 368 с.
2. Лямец В.И., Тевяшев А.Д. Системный анализ. – Харків: ХНУРЕ, 2004. – 448 с.
3. Павленко Л.А. Корпоративні інформаційні системи. Навчальний посібник. – Харків: Вид. ”ІНЖЕК”, ХНЕУ, 2005. – 260 с.
4. http://www.basegroup.ru/download/deductor/.

ИСПОЛЬЗОВАНИЕ АНАЛИТИЧЕСКОЙ ПЛАТФОРМЫ DEDUCTOR ПРИ ИЗУЧЕНИИ УЧЕБНОЙ ДИСЦИПЛИНЫ «ИНФОРМАЦИОННЫЕ АНАЛИТИЧЕСКИЕ СИСТЕМЫ»

Александрова В.А., *старший преподаватель Тверского филиала Московского государственного университета экономики, статистики и информатики, г. Тверь*

На всех уровнях управления современными компаниями актуальной является поддержка принятия своевременных и качественных управленческих решений. Использование в работе современных технологий поддержки принятия решений позволяет снизить риски, связанные с принятием необоснованных решений, корректировать тактику и стра-

тегию поведения на рынке в условиях быстро меняющейся ситуации. Внедрение информационно-аналитических систем помогает в решении этой задачи, обеспечивая возможность всестороннего анализа бизнеса, финансового планирования, прогнозирования, получения консолидированной отчетности.

Специалисты в различных областях (маркетинг, менеджмент, антикризисное управление, финансы и кредит и др.) должны имеет представление о содержании аналитической работы и владеть ее основным навыками. В связи с этим при подготовке специалистов в высших учебных заведениях актуальным является учебный курс «Информационные аналитические системы».

В ТФ МЭСИ в соответствии с учебным планом на изучение курса «Информационные аналитические системы» отводится 32 часа (16 часов – лекции, 16 часов – практические занятия). Дисциплина изучается студентами 3 и 4 курсов (в зависимости от специализации).

Согласно учебной программе целью изучения дисциплины является получение сведений о проблематике «автоматизации анализа информационной подготовки принятия управленческих решений с использованием современных инструментальных средств широкого применения и специализированных пакетов прикладных программ, освоение технологий оперативного и интеллектуального анализа данных, отражающих деятельность в различных предметных областях»[1].

В процессе изучения дисциплины студенты знакомятся с технологиями анализа, принципами и системами сбора и повышения качества исходных данных для анализа и последующего принятия решений, структурами информационных хранилищ, комплексами инструментальных средств, поддерживающих технологии анализа данных.

При изучении дисциплины «Информационные аналитические системы» используется академическая версия аналитической платформы *Deductor*, разработанная компанией BaseGroup Labs.

ТФ МЭСИ является партнером компании BaseGroup Labs в области образовательных программ.

Преподаватели вузов, обучающиеся по партнерской образовательной программе компании, имеют возможность использовать в учебном процессе отдельные методические материалы для лабораторных практикумов и дополнительные бизнес-задачи с исходными данными. Эти ма-

[1] Рабочая учебная программа дисциплины «Информационные аналитические системы», 2008.

териалы размещаются в специальном разделе для вузов-партнеров на сайте компании. Необходимо подчеркнуть, что это помогает в методической организации учебного курса, совершенствовании его теоретического и практического наполнения.

Также имеется возможность свободного доступа к размещенным на сайте компании теоретическим материалам по вопросам анализа данных. Примеры применения реализованных в Deductor аналитических технологий для решения реальных бизнес-задач приводятся в разделе «Практика анализа» (http://basegroup.ru/library/practice/). При изучении дисциплины «Информационные аналитические системы» студентам рекомендуется знакомиться с этими и другими материалами сайта компании (http://www.basegroup.ru/). В процессе изучения дисциплины на лекциях излагаются теоретические вопросы курса в соответствии с учебным планом. На практических занятиях студенты выполняют задания по созданию и наполнению хранилища данных, извлечению информации из хранилища; знакомятся с реализованными в аналитической платформе Deductor технологиями OLAP. Необходимо отметить, что Deductor отличается высококачественным дружественным интерфейсом и содержательной справочной системой, что играет немаловажную роль при обучении студентов. Следует также обратить внимание на встроенный в Deductor механизм визуализации, обеспечивающий удобное представление результатов анализа с точки зрения их интерпретации.

В рамках лекционного курса и практикума большое внимание уделяется технологиям Data Mining.

В процессе чтения лекций освещаются теоретические основы интеллектуального анализа данных и рассматриваются основные классы задач, решаемых методами Data Mining. При выполнении практических заданий студенты используют встроенные в *Deductor* инструментальные средства для решения задач классификации, кластеризации, поиска ассоциативных правил, прогнозирования временных рядов, прогнозирования на основе линейных и нейросетевых моделей. Причем в процессе решения конкретной задачи строятся несколько моделей с различными параметрами. Такой прием представляется методически обоснованным, поскольку важным этапом построения аналитических решений является оценка качества моделей. Студентами анализируются построенные модели, оценивается их качество с помощью таблиц сопряженности или диаграмм рассеяния, делаются аргументированные выводы по выбору модели. Следующим этапом является предложение и обоснование управленческих решений, которые можно принять, используя результаты анализа. В процессе выполнения заданий практикуется совместное обсуж-

дение наиболее интересных или сложных моментов по определенной тематике.

Среди тем, изучаемых в курсе «Информационные аналитические системы» актуальной, по моему мнению, является тема, посвященная вопросам качества данных, используемых для разработки аналитических решений. На лекции даются понятия качества данных и методы их оценки; обосновывается необходимость предварительной обработки исходных данных для улучшения их качества; рассматриваются инструменты предобработки, реализованные в *Deductor*, и особенности их использования. Это материалы в дальнейшем используется при выполнении практических заданий в процессе изучения различных разделов курса. Так, например, при решении задач прогнозирования строятся несколько нейросетевых моделей и оценивается их качество. Студенты проводят сравнительный анализ моделей. При этом имеется возможность убедиться в том, что наличие аномальных значений, дубликатов и противоречий приводит к построению моделей ненадлежащего качества; использование таких моделей, соответственно, оказывает влияние на снижение достоверности результатов анализа.

Практическое использование аналитической платформы *Deductor* при изучении дисциплины «Информационные аналитические системы» в сочетании с необходимым уровнем теоретической подготовки способствует приобретению навыков аналитической работы, что является важным для будущих специалистов.

ПРИМЕР ИСПОЛЬЗОВАНИЯ DEDUCTOR В ПОДГОТОВКЕ СПЕЦИАЛИСТОВ ПО ПРИКЛАДНОЙ ИНФОРМАТИКЕ В ИГЭУ

Баллод Б.А., *доцент Ивановского государственного энергетического университета, г. Иваново,* ***Муромкина А.В.,*** *к.м.н., врач-кардиолог,* ***Ковалев Д.Е.,*** *студент ИГЭУ*

Аналитическая платформа Deductor используется в Ивановском государственном энергетическом университете в учебном процессе подго-

товки специалистов по прикладной информатике с 2005 года. Студенты изучают основы использования платформы в специальном разделе (информационно-аналитические системы) курса «Информационные системы». В дальнейшем они активно используют полученные знания при выполнении курсовых и дипломных проектов, а также при выполнении творческих научно-исследовательских работ.

В качестве примера рассмотрена информационная система медицинской диагностики, а именно – прогнозирования восстановления ритма у больных с фибрилляцией предсердий.

К наиболее трудоемким задачам медицины относятся постановка диагноза и выбор курса лечения. Традиционно врачи решали эти задачи, полагаясь лишь на собственную интуицию и опыт. Сегодня в их арсенал все чаще входят способы, основанные на высоких технологиях и позволяющие обрабатывать большие потоки информации. Для этой цели применяются разнообразные модели, основанные на применении технологии интеллектуального анализа данных (Data Mining): деревья решений, нейронные сети, карты Кохонена и др.

Анализ вариабельности ритма сердца (ВРС) является неотъемлемой частью обследования кардиологических больных. В последние годы предпринимаются попытки оценки ВРС у больных с фибрилляцией предсердий (ФП). Разброс кардиоциклов объясняется отсутствием единого водителя ритма и особенностями атриовентрикулярного проведения при ФП.

Цель исследования: разработка интеллектуальной системы принятия решений о вероятности восстановления синусового ритма у больных с фибрилляцией предсердий.

На базе Data Mining легко создавать советующие системы для диагностики заболеваний, которые используют накопленные данные клинических исследований, автоматически выявляют значимые признаки и моделируют сложные зависимости между симптомами и заболеваниями.

Для проведения анализа была составлена обучающая выборка, содержащая значения факторов ВРС пациентов кардиологического диспансера, объемом более 100. После исключения выбросов и пропусков объем выборки, фрагмент которой представлен в табл. 1, составил N=86.

Таблица 1

Фрагмент обучающей выборки

ФИО	Группа	TP	VLF	LF	HF	RRNN	SDNNf	SDNNf \RRNNf
Пет***	A	11 379	1 611	2 094	7 675	704	131	0,186
Афе****	A	21 513	1 726	4 568	15 219	517	149	0,288
Ста*****	B	29 072	2 175	6 572	20 325	709	190	0,268
Фед****	A	32 528	5 393	9 532	17 602	903	195	0,216
Абр******	B	24 624	2 124	6 790	15 710	769	175	0,228
и т.д.								

Обозначения в таблице:

- Группа A – пароксизмальная форма нарушения ритма;
- Группа B – постоянная форма нарушения ритма;
- TP – общая мощность спектра ритмограммы;
- VLF – очень медленные волны;
- LF – медленные волны;
- HF – быстрые волны;
- RRNN – средняя продолжительность интервала между кардиоциклами;
- SDNNf – стандартное отклонение, квадратный корень из разброса интервалов RR;
- SDNNf \RRNNf – отношение стандартного отклонения к средней продолжительности интервала между кардиоциклами.

Деревья решений как вариант решения проблемы. Задачи классификации с большим успехом решаются одним из методов Data Mining – при помощи деревьев решений. Деревья решений – классификатор, полученный из обучающего множества, содержащего объекты и их характеристики, на основе обучения. Дерево состоит из листьев, указывающих на класс, и узлов, содержащих правила классификации. Оно может использоваться для классификации объектов, не вошедших в обучающее множество. Получаемая модель – это способ представления правил в иерархической, последовательной структуре, где каждому объекту соответствует единственный узел, дающий решение.

Построение дерева решений по алгоритму C4.5 реализовано соответствующим модулем аналитической платформы *Deductor*, диалоговое окно которого показано на рис. 1.

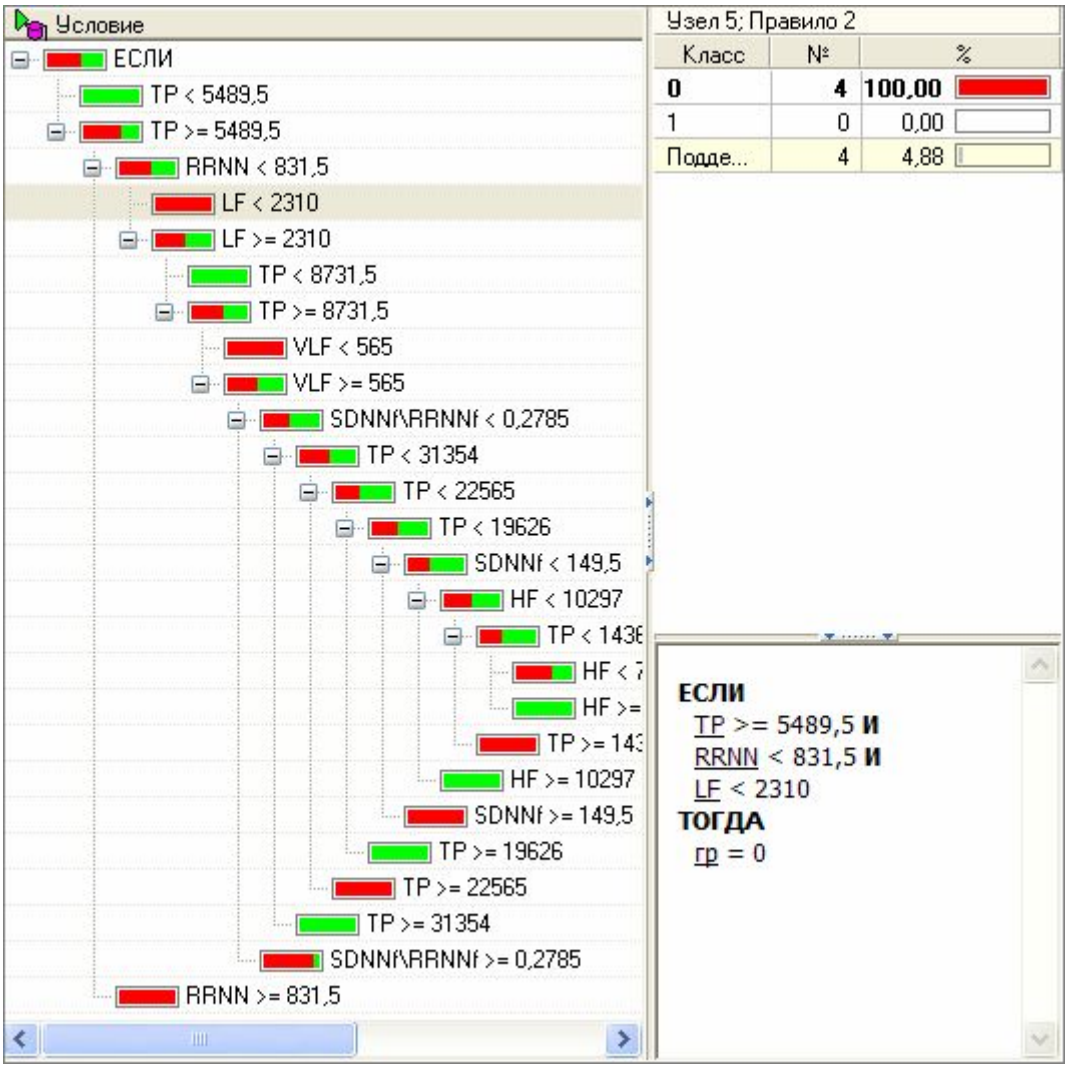


Рис. 1. Пример дерева решений

Полученную модель можно использовать при определении класса заболевания, к которому следует отнести вновь поступившего больного.

В алгоритме C4.5 каждый узел дерева решений может иметь нескольких потомков. Для выбора оптимального разделяющего правила используется функция оценки качества разбиения. Оценка качества разбиения формализована в индексе *Gini*. Если набор данных *T* содержит данные *n* классов, тогда индекс *Gini* определяется как:

$$Gini(T) = 1 - \sum_{i=1}^{n} p_i^2 ,$$

где p_i – вероятность (относительная частота) класса *i* в *T*.

На рис. 2 представлены ошибки классификации на обучающей выборке объемом *N*=86. Количество ошибок в группе *A* = 5, в группе *B* = 3, что составляет менее 10% от числа диагностируемых. Такой результат вполне пригоден для практического использования.

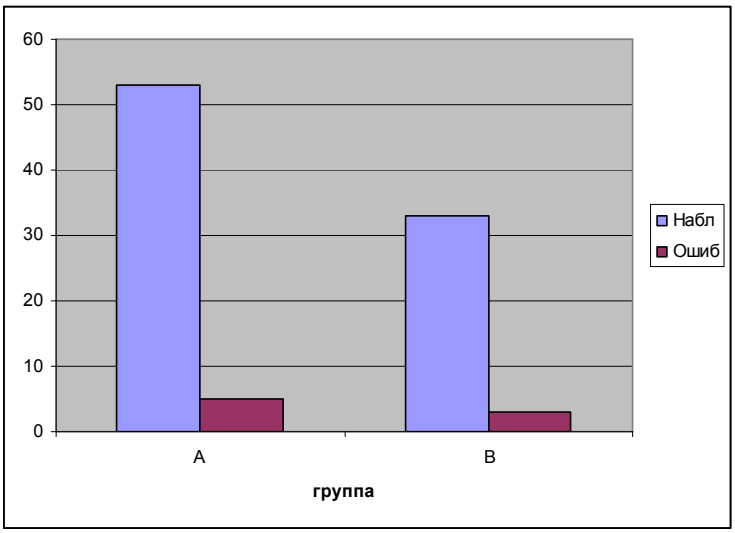


Рис. 2. Результат тестирования модели

Для практического применения модель дерева решений была реализована средствами *MS Excel* популярного офисного пакета MS Office. Интерфейс программы приведен на рис. 3.

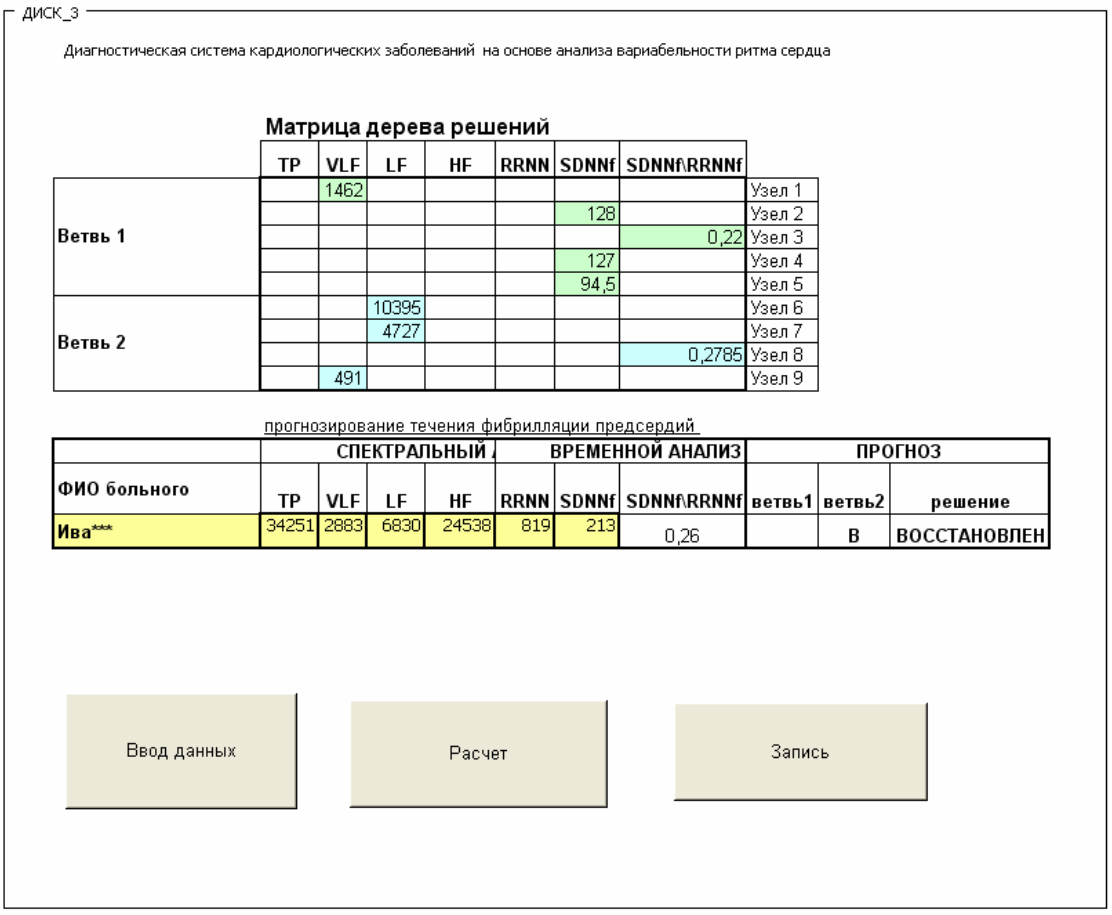


Рис. 3. Диалоговое окно программы «ДИСК_3»

Технология диагностирования больных в разработанной информационной системе предусматривает следующие этапы.

Этап 1. С помощью программы «Поли-спектр» производится запись и первичный анализ кардиограммы больного (рис. 4).

Отделение: кардио - 3

Пациент: Ива*****, 44 года, 180 см, 80,0 кг, ИМТ: 24,7 (норма)

Предварительный диагноз: ГБ 2ст. Персист. ФП (перед ЭИТ).

Тип обследования: ВРС от 18.08.2009

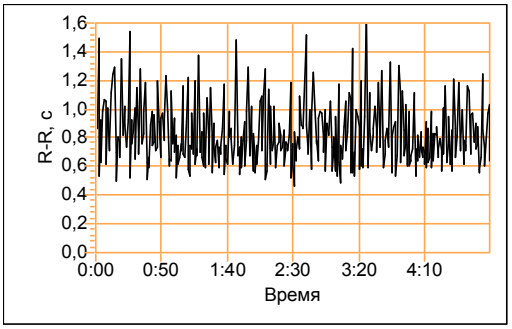
Прибор: Поли-Спектр-8/ЕХ

Фоновая запись

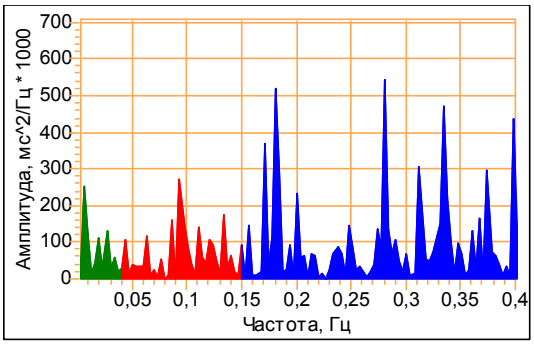
Положение пациента при записи: лежа. Продолжительность записи сердечного ритма 5 мин.(300 с). Зарегистрировано комплексов QRS: 366.

Ритм ФП. Средняя ЧСС: 73 уд./мин. – нормальная частота сердечного ритма.

Ритмограмма



Спектрограмма



Показатели

временного анализа		спектрального анализа		Структура спектра	
Параметр	Значение	Параметр	Значение	Параметр	Значение
R-R min, мс	465	TP, мс^2	34251	%VLF	8,42
R-R max, мс	1667	VLF, мс^2	2883	%LF	19,9
RRNN, мс	819	LF, мс^2	6830	%HF	71,6
SDNN, мс	213	HF, мс^2	24538		
RMSSD, мс	316	LF norm, n.u.	21,8		
pNN50, %	85,8	HF norm, n.u.	78,2		
CV, %	26,1	LF/HF	0,278		

Рис. 4. Фрагмент протокола анализ кардиограммы больного

Этап 2. Данные предварительного анализа заносятся в диалоговую форму программы «ДИСК_3», представленную на рис. 3.

Этап 3. После выполнения расчета в окне «прогноз» выводится решение интеллектуальной системы: «ВОССТАНОВЛЕН» (или «АРИТМИЯ») (рис. 3), что означает прогнозируемую оценку восстановления ритма сердца после соответствующего лечения. Такое решение принято, потому что в процессе диагностирования по результатам предварительного анализа больной был отнесен к группе «В» – с пароксизмальной формой нарушения ритма сердца, т.е. с исходом в восстановление ритма.

Одобренное врачом решение заносится в базу данных, которая служит обучающей выборкой для дальнейшей корректировки модели Дерева с помощью аналитической платформы *Deductor*.

Этап 4. На основании предварительного анализа ВРС и решения интеллектуальной системы врач делает заключение о вероятности восстановления синусового ритма у конкретного пациента с ФП.

Выводы.

1. Методом интеллектуального анализа данных (Data Mining) на основе алгоритма деревья решений аналитической платформы Deductor построена модель диагностирования формы ФП (пароксизмальная или постоянная) и прогнозирования восстановления ритма у этих больных.

2. Модель реализована в виде программного продукта ДИСК_З, и может использоваться в виде интеллектуального помощника врача-кардиолога.

Литература

1. Паклин Н.Б., Орешков А.И. Бизнес-аналитика: от данных к знаниям(+CD). – СПб.: Питер, 2009.
2. Михайлов В.М. Вариабельность ритма сердца. – Иваново: НейроСофт, 2000.
3. Баллод Б.А., Чайкин М.О. Разведка данных в среде Deductor. – Иваново: ГОУ ВПО ИГЭУ, 2008.

ПРОБЛЕМЫ ПРИМЕНЕНИЯ МЕТОДОВ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА ДАННЫХ В СИСТЕМАХ ПОДДЕРЖКИ ПРИНЯТИЯ РЕШЕНИЙ

Болотова Л.С., *профессор,* ***Кузнецов С.Н.****, студент, МИРЭА,* ***Демина Н.Н.****, студент, РГУИТП*

Наш интерес к интеллектуальному анализу данных (ИАД) начался со знакомства с монографией по бизнес-аналитике [1], выпущенной в середине 2009 года.

В МИРЭА читаются достаточно серьезные курсы по теории принятия решений и системам искусственного интеллекта. При этом акцентируется внимание на проблемы разработки Баз Знаний (БЗ) для систем поддержки принятия решений (СППР). В основу курсов положены авторские разработки – метод ситуационного анализа и проектирования БЗ или модели предметной области (МПрО) произвольной природы [2, 3, 4]. На основе метода разработана сквозная технология, программные средства, обеспечивающие построение БЗ СППР и их реализацию. Технология предполагает, что в результате направленной работы с экспертами проектируется концептуальная модель предметной области, которая затем переводится в объектно-ориентированное представление, так называемую понятийно-объектную модель. Она является основой для автоматической генерации БЗ в любом из нужных представлений: продукционном, фреймовом, гибридном – в зависимости от используемой для вывода решений инструментальной системы логического вывода.

Как известно, СППР является вершиной айсберга любой автоматизированной системы управления, информационно и программно связанной через Хранилище Данных (ХД) с системой Баз Данных (БД), обеспечивающих функционирование триады: *Данные от объекта управления и внешнего мира → ХД → СППР*.

Однако в настоящее время эта триада разорвана на две практически независимые части: Данные и система их обработки, включая интеллектуальный анализ данных и, собственно, СППР, тоже со своими информационными и программными средствами. На наш взгляд, такая ситуация является тормозом для развития как теоретических, так и инструментальных средств этих частей, как единого целого.

Именно этот аспект стал основой нашего интереса к методам ИАД. Мы занимались СППР без опоры на реальные потоки данных, а

для БЗ не предусматривалась возможность автоматического её пополнения (обучения) за счет скрытых знаний, выявляемых в результате их анализа.

В этой связи мы задались рядом вопросов, на которые пытаемся сегодня ответить:

- каким образом формировался документооборот предприятий и, соответственно, состав показателей и БД (стихийно – исторически или на основе сознательного (целевого) проектирования);
- как руководители распознают проблемные ситуации на предприятиях;
- в какой мере потребности ЛПР удовлетворяются существующей отчетностью;
- какой должна быть система показателей и отчетности предприятия, для того, чтобы они наилучшим образом соответствовали требованиям руководителей разных уровней;
- какие запросы могут поступать от системы поддержки принятия решения;
- как должно быть организовано ХД, исходя из потребностей ЛПР.

Известно, что ХД – разновидность систем хранения, ориентированная на поддержку процесса анализа данных, обеспечивающая целостность, непротиворечивость, и хронологию данных, а так же высокую скорость выполнения аналитических запросов [1]. Хранилище данных максимально обеспечивает возможность обработки и анализа данных, выявление закономерностей и правил, которые в свою очередь являются необходимыми критериями точности работы СППР.

Основой хранилища данных является некий процесс, модель которого строится в хранилище (полный спектр необходимых объектов, субъектов, их свойств, действий и компонентов действий). Здесь возник ряд вопросов по поводу технологии проектирования самих ХД:

- каким образом формируется структура ХД: как набираются измерения, атрибуты, процессы;
- как технология проектирования ХД поддерживается теоретически и т.д.

Мы пришли к выводу, что общий подход к определению структуры ХД – числа процессов, связей между процессами, числа измерений и атрибутов, отсутствует, что данная проблема теоретически разработана слабо и в данный момент построение хранилища данных можно отнести к искусству инженера-аналитика.

Мы начали с того, что в курс по системам искусственного интеллекта, был введен раздел по ИАД, включающий методы Data Mining, в частности, методы кластеризации, классификации и ассоциации, а также метод ДСМ – автоматического порождения гипотез, разработанный Финном В.К. [5, 6, 7]. Соответственно, были разработаны методики и варианты проведения лабораторных работ по этим методам. А в качестве обучающих выборок из БД мы предлагали темы курсовых работ, которые студенты выполняли в рамках курса по теории принятия решений. При этом они самостоятельно выбирали предметную область, определяли типы проблемных ситуаций, требующих принятия решений, строили модель предметной области по нашей методике с помощью авторских программных средств и, в конце концов, создавали БЗ и прототип экспертной системы. Им предлагалось, исходя из концептуальной модели и БЗ, спроектировать основную систему показателей и необходимых БД для обеспечения полного цикла функционирования СППР, т.е. всей триады: *Данные от объекта управления и внешнего мира → ХД → СППР*. Во втором варианте, который был также поставлен и проверен на лабораторных работах, состоял в том, что мы предлагали выполнить распознавание и классификацию конкретных объектов из обучающей выборки путём построения прототипа нечеткой экспертной системы, а затем, используя методы построения дерева решений и ассоциации проверить их достоверность. В итоге мы стремились к формированию у студентов навыков для оценки применимости различных методов ИАД и сознательного выбора наиболее подходящих в каждом конкретном случае. Кроме того, такой подход обеспечивает формирование концептуального мышления и системного взгляда на проблему у студентов.

С использованием методов ИАД на материале лаборатории эпидемиологической кибернетики НИИ им. Гамалея Н.Ф. выполнен проект по разработке информационного и программного обеспечения СППР для противодействия распространению эпидемий посредством авиапассажирского трафика. Основная идея проекта состояла в сборе, подготовке и обработке оперативных данных по авиапассажирскому трафику на основе расписаний полетов самолетов по всем столицам мира пока только на основе расписания полетов США[1]. Мы считаем эту работу очень интересным опытом использования методов ИАД в составе СППР. Начатые работы мы предполагаем продолжить и в следующем учебном году. Нужно добавить, что такого рода работы относятся к научно-исследовательским, и надеемся, что это поможет не только обогатить

[1] См. соответствующую статью авторов *Болотова Л.С., Боев Б.В., Демина Н.Н.* в этом сборнике

учебный процесс, но и воспитать молодых ученых с современным взглядом на важные проблемы человечества в целом.

Опыт работы одного года позволил нам уже сформулировать пожелания к разработчикам аналитической платформы *Deductor*:

- в академической версии желательно расширить число возможных источников данных, текстового формата недостаточно;
- в документации желательно расширить теоретические основы и практические возможности по построению многомерных хранилищ данных;
- неплохо бы продумать создание единого обменного фонда БД для решения различных типов задач: медицины, сельского хозяйства, производства, а не только по бизнесу и управлению.

Такой обмен между вузами мог бы существенно расширить и стимулировать работы по ИАД. И ещё одно пожелание. Может быть, компания возьмет на себя опеку и помощь вузам в становлении научно-образовательных центров по бизнес-аналитике, т.е. созданию центров, ведущих коммерческую деятельность, помогающих в трудоустройстве выпускников. Очевидно, для этого должны быть разработаны соответствующие правовые основы.

Литература

1. Паклин Н.Б., Орешков В.И. Бизнес-аналитика: от данных к знаниям. – СПб.: Питер, 2009. – 624 с.
2. Болотова Л.С. Модели представления знаний в системах искусственного интеллекта. Ч. 1. Теоретические основы искусственного интеллекта и формальные модели: Учеб. пособие. – М.: РГУИТП, 2009. – 129 с.
3. Болотова Л.С. Модели представления знаний в системах искусственного интеллекта. Ч. 2. Неформальные модели: Учеб. пособие. – М.: РГУИТП, 2009. – 130 с.
4. Болотова Л.С., Смольянинова В.А., Смирнов С.С. Концептуальное проектирование модели предметной области при помощи программных систем разработки баз знаний для интеллектуальных систем поддержки принятия решений // Наукоемкие технологии. – М.: Радиотехника, 2009. – Т.10. – №8. – С. 23–28.
5. Арский Ю.М., Финн В.К. Принципы конструирования интеллектуальных систем // Информационные технологии и вычислительные системы. – 2008. – № 4. – С. 95–127.
6. Аншаков О.М. Об одном подходе к порождению гипотез в ДСМ-методе // Десятая национальная конференция по искусственному интеллекту с международным участием КИИ-2006: Тр. конф. – М: Физматлит, 2006.

7. Получение знаний для формирования информационных образовательных ресурсов: Учеб. пособие для вузов / А.Д. Иванников [и др.]. – М.: Информика, 2008. – 438 с.

ОБ ИНТЕГРАЦИИ DEDUCTOR С ДРУГИМИ ИНФОРМАЦИОННЫМИ СИСТЕМАМИ

Носков В.В., *доцент,* ***Прокопенко Н.Ю.,*** *доцент,,* ***Рабынина О.В.,*** *студент, Нижегородский государственный архитектурно-строительный университет, г. Нижний Новгород*

Современные тенденции создания интегрированных автоматизированных систем включают разные аспекты интеграции – интеграцию данных, технологий и технических средств.

Технологии интеграции различных систем обработки данных были реализованы еще в эпоху дисковых операционных систем. В качестве примера можно назвать такие интегрированные пакеты (распространяемые еще в СССР), как *Open Access*, *Framework* и др. Классическим примером современных интегрированных систем являются распространенные пакеты *MS Office* и *Open Office*.

Интеграция данных (Data integration) – процесс объединения данных из различных источников для получения их согласованного представления. В широком смысле – это процесс организации регулярного обмена данными между различными информационными системами (ИС) предприятия или организации. Проблема интеграции данных является неотъемлемым аспектом проблематики развития информационной инфраструктуры и следуемого объекта (предприятия или организации).

В данной работе реализована интеграция информационных систем, имеющих разные специализации, но включающих в свои процессы обработки данных идеологию реляционных структур данных. В качестве базовых информационных систем используются СУБД *MS Access* [2], аналитическая платформа *Deductor Academic* [1] и геоинформационная система *MapInfo* [3]. На базе этих трех систем строится информационно-аналитическая система (ИАС) анализа детской смертности по Нижегородскому региону. Основная часть обработки информации выполняется в приложении СУБД *MS Access*, аналитическая обработка в Deductor Academic, а графическая интерпретация результатов в *MapInfo*.

Разработанная интегрированная ИАС предназначена для комплексного информационно-аналитического обеспечения информационной и методической поддержки сотрудников организационно-методического

отдела Нижегородской областной детской клинической больницы (НОДКБ), для повышения качества и оперативности принятия решений по проблемам детской смертности. Данную систему предполагается использовать в НОДКБ и её территориальных органах.

Цели создания системы:

- повышение полноты и достоверности электронных баз данных и данных статистической отчётности;
- уменьшение трудозатрат, сокращение времени сбора и подготовки отчетных и аналитических материалов детской областной больницей;
- сокращение объема бумажного документооборота и улучшение координации совместной работы отделений НОДКБ и органов государственной статистики (ОГС) при подготовке аналитических материалов;
- совершенствование процедур оперативного анализа значений показателей детской смертности Нижегородского региона, а также моделирования и прогнозирования показателей смертности и рождаемости.

Объектом автоматизации является система информационного взаимодействия между подразделениями НОДКБ и ОГС. Информационное взаимодействие осуществляется посредством предоставления ОГС статистических данных НОДКБ, а также через получение оперативных данных о детской смертности организационно-методическим отделом НОДКБ, данных из других отделений больницы и аккумуляции их в Централизованном хранилище данных.

В качестве исходной информации для ИАС были взяты данные за период 1998-2008 гг. Данные представляют собой статистику рождаемости за последние 10 лет и данные о смертности детей в возрасте от 0 до 17 лет.

На рис. 1 приводится схема модели, описывающая потоки данных в ИАС. Диаграмма потоков данных показывает внешние по отношению к системе источники данных, которые принимают информацию от систем, а также идентифицируют хранилища данных, к которым осуществляется доступ системы.

Анализ данной модели приводит к следующему составу подсистем АИС (рис. 2):

- подсистема сбора и хранения данных;
- подсистема очистки и подготовки данных к анализу и построению аналитических моделей;
- подсистема моделирования и прогнозирования;

- подсистема аналитической отчетности и географического отображения данных.

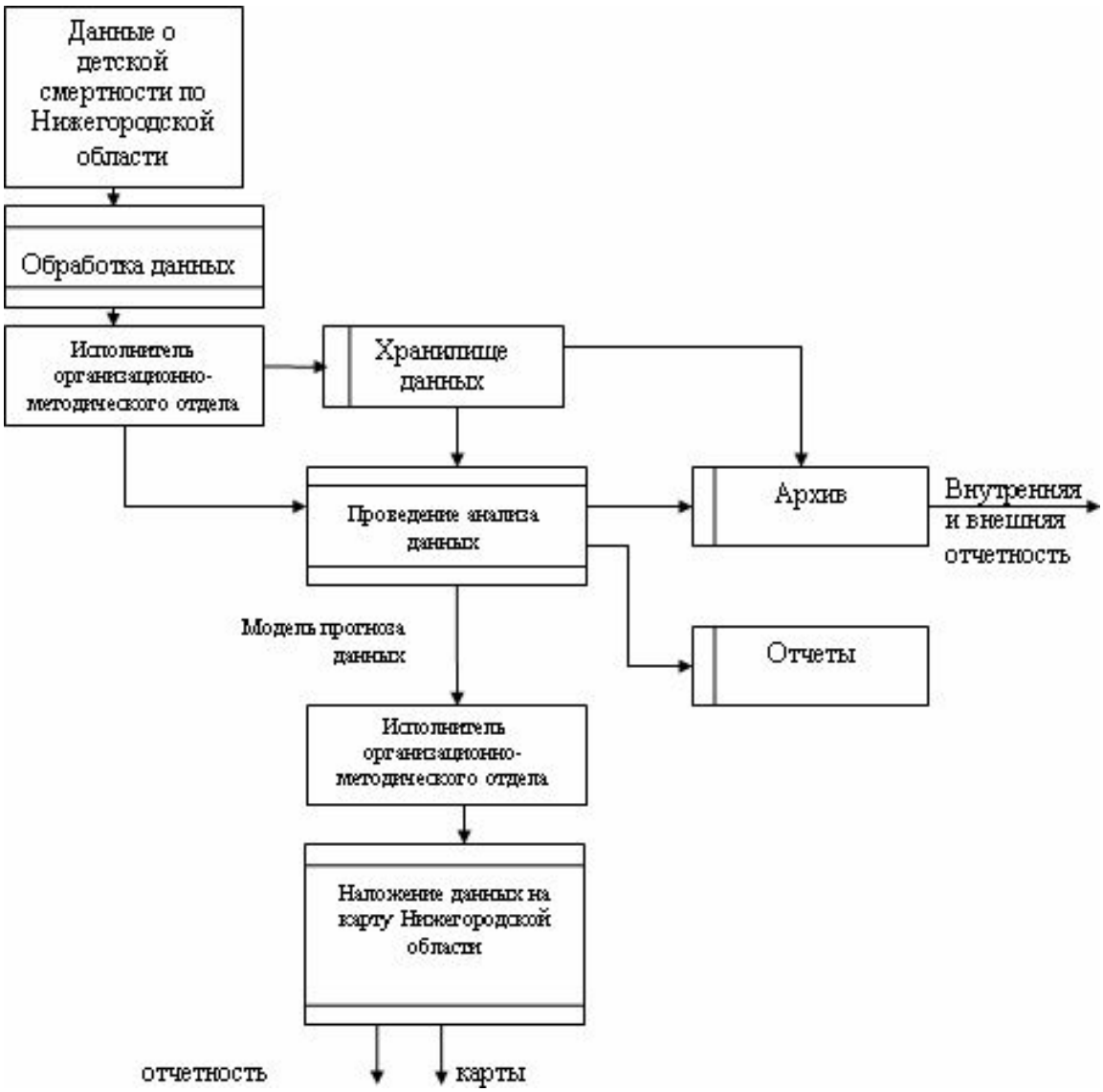


Рис. 1. Диаграмма потоков данных

Рис. 2. Подсистемы ИАС

ИАС предполагает повышение эффективности исполнения рабочих процессов организационно-методическим отделом путем сокращения

непроизводительных и дублирующих операций, операций, выполняемых «вручную», оптимизации информационного взаимодействия участников процессов.

ИАС предназначена для решения следующих основных задач:

- сбор, интеграция, структурирование, хранение и ретроспективное накопление информации о детской смертности в хранилище данных, содержащем общую статистику о случаях летальных исходов среди детей трех возрастных групп (младенческая, детская и подростковая) по районам нижегородской области за несколько лет;
- обработка, группировка данных и построение аналитической отчетности;
- комплексный анализ проблемных вопросов с использованием математических, интеллектуальных методов и современных информационных технологий, получение агрегированных текущих и прогнозных оценок, как по отдельным причинам, так и по возрастным группам;
- нанесение полученных данных на карту Нижнего Новгорода и Нижегородской области;
- комплексная оценка проблемы детской смертности в Нижегородской области;
- информационно-аналитическое и информационно-справочное обеспечение руководителей организационно-методического отделения и руководителей смежных отделений детской областной больницы.

Подсистема сбора и хранения информации реализуется средствами СУБД *MS Access* и *Deductor Academic*. Существуют две идеологии хранения данных: базы данных и хранилища данных. База данных предназначена для хранения оперативных и статистических данных, которые используются в процессе подготовки аналитических отчетов. Данные хранятся в виде коллекции таблиц, где общие поля используются для связей. Эта часть системы реализована в СУБД *MS ACCESS*. Хранилище данных – разновидность систем хранения. Это специально организованная база данных, ориентированная на решение задач анализа данных и поддержки принятия решений, обеспечивающая максимально быстрый и удобный доступ к информации, целостность, непротиворечивость и хронологию данных, а также высокую скорость выполнения аналитических запросов.

Назначение хранилища данных – своевременно обеспечить аналитика всей информацией, необходимой для проведения анализа, построе-

ния моделей и принятия решений. Хранилище данных реализовано на базе *Deductor Academic*.

Цель хранилища данных не анализ данных, а подготовка данных для анализа и их консолидация. Поэтому хранилище данных входит и в подсистему очистки и подготовки данных к анализу. Эта подсистема предназначена для аудита данных, подготовки данных к анализу и построению аналитических моделей. Аудит данных включает в себя: проверку и устранение дубликатов и противоречий, обработку пропусков, выявление выбросов и фильтрацию. Качество данных, которые собираются и консолидируются для анализа из различных источников, является одной из самых больших проблем аналитических технологий. Недостаточное внимание к проблеме качества данных способно свести на нет все преимущества самых современных и мощных методов анализа, все усилия аналитиков и экспертов по созданию аналитических решений. С целью повышения качества данных используется комплекс методов и алгоритмов, получивших название «очистка данных».

Подсистема моделирования и прогнозирования реализуется в инструментальной среде аналитической платформы *Deductor Academic*, который включает в себя универсальные мастера и обработчики, обеспечивающие импорт, обработку, построение моделей, визуализацию и экспорт данных. Подсистема моделирования и прогнозирования предназначена для поиска функциональных и логических закономерностей в накопленной медико-статистической информации, для построения моделей и правил, которые объясняют найденные закономерности и прогнозируют изменение медико-статистических показателей, развитие медико-демографических процессов.

Автоматизированный процесс исследования тенденций, моделей и взаимосвязей в данных включает в себя применение статистических методов и методов искусственного интеллекта для анализа исходной информации и выявления скрытых закономерностей, которые не могут быть обнаружены непосредственно или на интуитивном уровне.

На этапе анализа проверяется корреляция между факторами регион, причина, возраст неонатальный, группа возрастная, местность, место смерти, масса тела. Расчет парной корреляции между факторами, которые могут быть ранжированы (упорядочены) производят с помощью рангового коэффициента Спирмена. Для пары признаков причина и возраст неонатальный, причина и возрастная группа было доказано существование корреляционной зависимости между факторами. В работе был также рассчитан коэффициент множественной корреляции по Кендаллу (коэффициент конкордации) и был сделан вывод, что зависимость между всеми факторами в совокупности незначительная.

Подсистема прогнозирования ИАС строит три модели: нейросетевой прогноз временного ряда показателей детской смертности, прогнозирование временного ряда с использованием парциальной обработки и скользящего окна, прогнозирование временного ряда на основе линейной регрессии. Производится сравнений моделей и делаются выводы о том, с помощью какой модели получается наиболее точный прогноз. Лучшей оказалась модель прогноза временного ряда с использованием парциальной обработки и скользящего окна, так как полученная разница между фактическим значением детской смертности за 2009 год и полученным прогнозным значением всего 1 %. Аналогичным образом для показателей рождаемости было построено три модели прогноза. В случае прогнозирования рождаемости лучшей является первая модель – нейросетевой прогноз временного ряда (разница между прогнозным и фактическим значением – менее 1 %).

Для представления результатов анализа и прогнозирования для конечных пользователей используется подсистема аналитической отчетности и географического отображения данных. Целью создания подсистемы отображения данных является визуализация данных: получение аналитических отчетов OLAP-средствами и карт с помощью геоинформационных систем (ГИС).

Было создано ряд отчетов средствами *Access* (по требованию заказчика) и в *Deductor*. Аналитическая отчетность в Deductor обеспечивает быстрый доступ к результатам анализа, не требуя от пользователя навыков анализа данных и работы в системе. При работе с отчетами пользователь не видит сценарий анализа данных, ему доступны только конечные результаты (выдержки) из работы аналитика. Отчеты построены в виде древовидного иерархического списка, каждым узлом которого является отдельный отчет или папка, содержащая несколько отчетов. Каждый узел дерева отчетности связан со своим узлом в дереве сценария. Для каждого отчета были построены OLAP-кубы, а также настроены другие способы отображения: гистограммы, кросс-таблицы, кросс-диаграммы.

Как средство визуализации в подсистеме отображения данных используются также геоинформационные системы (ГИС) (рис. 3). ГИС предназначены для сбора, хранения, анализа и графической визуализации пространственных данных и связанной с ними информации о представленных в ГИС объектах. ГИС предоставляет набор средств создания и объединения баз данных с возможностями их географического анализа и наглядной визуализации в виде различных карт, графиков, диаграмм, прямой привязке друг к другу всех атрибутивных и графических данных. Геоинформационные системы – многофункциональные средства анализа

сведенных воедино табличных, текстовых и картографических данных. Подсистема географического отображения реализована средствами геоинформационной системы *MapInfo*. ГИС расширяет возможности системы и позволяет упростить аналитические работы с координатно-привязанной информацией.

Главная форма ИАС, реализованная средствами СУБД *MS Access*, выполняет функцию интеграции приложений и является главной панелью управления, представляя запуск систем MS *Access*, *Deductor*, *MapInfo*, а также работу с подсистемами.

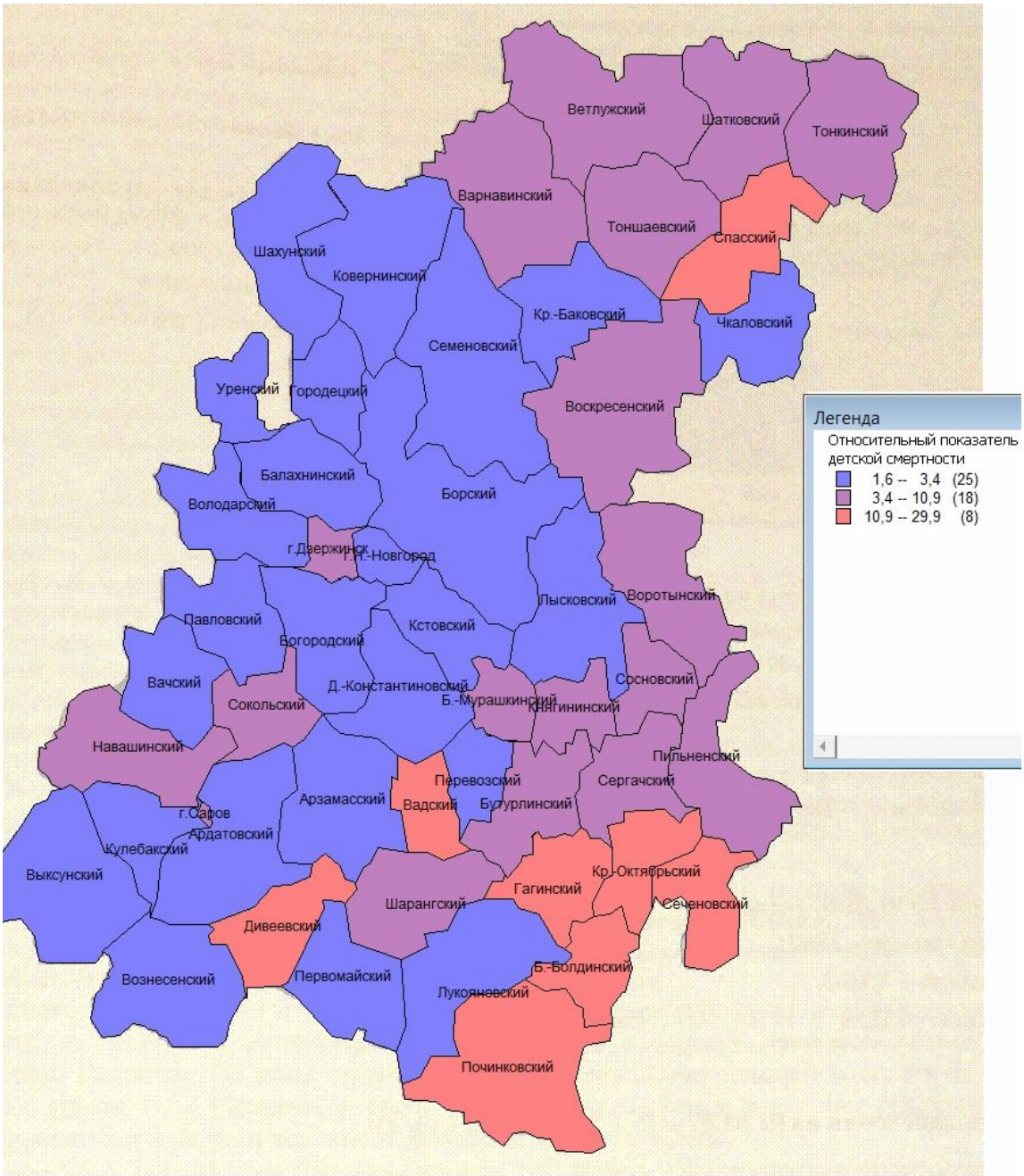


Рис. 3. Тематическая карта – относительные показатели детской смертности по регионам Нижегородской области

Литература

1. Карпузова В.И. Системы поддержки принятия решений на базе решений аналитической платформы Deductor Studio Academic 5.1: учеб. Пособие / Карпузова В.И., Скрипченко Э.Н., Стратонович Ю.Р., Чернышева К.В.. – М.: ФГОУ ВПО РГАУ – МСХА, 2010 – 75 с.
2. Джон Вескас. Эффективная работа с Microsoft Access 2003. – СПб.: Питер, 2005. – 1168 с.
3. Рекомендации по пользованию системой MapInfo 5.0 – http://www.map-info.ru.

СЕКЦИЯ
«АКТУАЛЬНЫЕ ЗАДАЧИ БИЗНЕС-АНАЛИТИКИ И ИХ РЕШЕНИЕ АЛГОРИТМАМИ DATA MINING»

МЕТОДЫ DATA MINING В СИСТЕМАХ КОНТРОЛЯ СОСТОЯНИЯ СЛОЖНЫХ ТЕХНИЧЕСКИХ СИСТЕМ

***Потюпкин А.Ю.**, начальник кафедры Военной академии РВСН имени Петра Великого, г. Москва*

Сложные системы сегодня представляют уже не только теоретический интерес. Сложными являются практически любые системы, способные к непредсказуемому для наблюдателя поведению, обладающие тем, что принято называть скрытыми закономерностями поведения. Сложные системы нелинейны, они способны к необратимому качественному развитию, обладают свойствами самоорганизации и саморазвития. Такая система для наблюдателя всегда остаётся «вещью в себе». Предсказать её поведение абсолютно точно невозможно. Как правило, к таким системам относят эргатические системы, в которых очень велико влияние человеческого фактора. Однако и технические системы, сконструированные как «простые», зачастую ведут себя как сложные системы. Помимо «больших» технических систем, в которых число подсистем очень велико, а состав разнороден, сложными являются и относительно небольшие системы, обладающие встроенными средствами ЭВТ с развитым программным обеспечением. Такие системы, функционируя большей частью по известным алгоритмам, и являясь для наблюдателя достаточно «простым» объектом, вдруг начинают вести себя непредсказуемо и быстро становятся «сложными». На практике всякая нештатная ситуация может быть отнесена к проявлениям сложности системы.

Управление сложной системой, как и любой системой, предполагает контроль её состояния, который по своей сути является обратной связью в контуре управления. В общем виде процесс определения технического состояния состоит из трех последовательных и взаимосвязанных этапов:

1. накопления и обобщения априорной информации о системе;
2. получения контрольно-диагностической информации о системе;
3. разработки и применения правил принятия решений о техническом состоянии.

Структура процесса контроля в целом представлена на рис. 1.

Рис. 1. Структура процесса контроля

Конечной целью этапа накопления и анализа априорной информации о системе является создание своего рода среднего эталонного образа каждого класса технического состояния.

Этап получения контрольно-диагностической информации о системе имеет целью получение апостериорной информации посредством производства измерений контролируемых параметров и представления результатов измерений на языке словаря признаков контроля.

Заключительным этапом является этап применения разработанных на первом этапе правил принятия решения о техническом состоянии, или этап «распознавания» образов технических состояний в результате композиции априорной и апостериорной информации о системе. Этот этап является наиболее сложным и ответственным, так как его результатом

является заключение о техническом состоянии системы. Данный этап рассматривают как этап анализа контрольно-диагностической информации с целью получения заключений о состоянии системы.

В штатных ситуациях, то есть ситуациях, предусмотренных эксплуатационно-технической документацией, задача анализа состояния является обычной задачей классификации. Однако в нештатных ситуациях, когда налицо проявление признаков сложности системы – нелинейности, проявления скрытых тенденций поведения, возникает необходимость решения задач выделения и распознавания новых состояний, определения неявных зависимостей между параметрами системы, выявления наиболее влияющих факторов, формирования дополнительных признаков контроля и ряда других. Такие задачи могут быть отнесены к задачам идентификации или в ряде случаев кластерного анализа, задачам поиска ассоциаций, секвенционального анализа, задачам регрессионного и факторного анализа.

В существующих контурах управления техническими системами задачи контроля и анализа полученной контрольно-диагностической информации в основном автоматизированы, например, процессы сбора и обработки информации с целью подготовки исходных данных для анализа состояния являются практически полностью автоматизированными. Однако применяемое программно-алгоритмическое обеспечение анализа обладает ограниченными возможностями – как правило, оно рассчитано на штатные условия функционирования, при этом анализу подвергаются только явные тенденции поведения системы, объект контроля рассматривается только как «простая» система. В нештатных ситуациях программные комплексы анализа не позволяют решать задачи контроля, и анализ состояния проводится полностью силами специалистов – анализаторов. В связи с этим говорить об автоматизации анализа сложных технических систем не представляется возможным – автоматизация анализа на практике скорее исключение, чем правило, что свидетельствует о наличии ряда существенных противоречий, порождающих проблемные вопросы.

В общем случае автоматизация предполагает создание модели объекта на базе формального аппарата моделирования, обеспечивающего адекватность модели, и её реализацию в виде совокупности алгоритмов и программ. В связи с этим, проблема автоматизации не существует изолированно, она тесным образом увязана с вопросами обеспечения адекватности моделей и выбора формально-математического аппарата. При решении задачи анализа состояния сложных систем модель должна отражать сложность объекта анализа, а формальный аппарат обеспечивать корректность её решения как обратной задачи. В силу этого, проблема

автоматизации анализа должна рассматриваться комплексно наряду с проблемами обеспечения корректности анализа и моделирования сложных систем.

Ситуация осложняется и высокими требованиями по оперативности управления, в том числе вплоть до реального масштаба времени, большими объёмами контрольно-диагностической информации – до десятков тысяч одновременно измеряемых параметров, высокими требованиями по надёжности программного обеспечения.

Можно констатировать, что анализ сложной технической системы представляет собой крайне сложную задачу и может быть реализован только на принципах динамического анализа с привлечением разнородных источников информации, а сама система анализа должна представлять собой интеллектуальную систему. В силу этого проблема анализа сложной системы может быть разрешена путем создания системы анализа с логикой, близкой к человеческому мышлению, способной к восприятию и обработке разнородных источников информации, как синтаксической так и семантической. Такая система должна быть устойчивой по отношению к влиянию неопределенных факторов для обеспечения корректности анализа, что достигается целенаправленным вмешательством в процесс не только анализа, но и сбора и первичной обработки для подготовки исходных данных.

Кроме того, не следует упускать из виду проявления так называемой глобальной проблемы технической кибернетики: «проецирования теоретически оптимальных методов на ограниченные возможности вычислительной техники». От того, каким образом будет реализована на практике система анализа, являющаяся средством специалиста-анализатора, зависит и выполнение требований корректности и интеллектуальности анализа. В частности, немаловажную роль играет обеспечение дружественного интерфейса «человек – машина», основой которого должен послужить в первую очередь естественный язык. В рамках такой системы будет реализовываться технология анализа, как целенаправленная совокупность процессов подготовки исходных данных, анализа (идентификации состояния) и выработки заключений о состоянии анализируемого объекта.

В связи с этим актуальной является разработка универсальных аналитических платформ анализа, ориентированных в первую очередь на технические приложения. Такие платформы должны быть сертифицированы, позволять оперативно обрабатывать большие объёмы количественной и качественной информации, обеспечивать высокую достоверность выдаваемых результатов.

Представляется, что методологической основой таких комплексов

могут послужить методы Data Mining, которые, во-первых, позволяют построить адекватную информационную модель сложной технической системы, а во-вторых, предоставляют аналитику разнообразный инструментарий для решения многих задач анализа. На сегодняшний момент разработан и широко известен ряд программных приложений, реализующих методы Data Mining. Однако главным их недостатком является узконаправленность и ориентированность в основном на сферу экономики и торговли, где сами методы получили наибольшее распространение.

К числу задач контроля и анализа состояния сложных технических систем, решаемых с помощью методов Data Mining, можно отнести следующие.

1. *Задача классификации* – определение класса объекта по его характеристикам – как задача определения одного из штатных технических состояний системы.
2. *Задача регрессии* – определение по известным характеристикам объекта некоторых его параметров – как задача факторного анализа.
3. *Задача поиска ассоциативных правил* – нахождение частых зависимостей (ассоциаций) между объектами или событиями – как задача определения неявных зависимостей между параметрами системы, задача секвенционального анализа.
4. *Задача кластеризации* – поиск независимых групп (кластеров) и их характеристик во всем множестве анализируемых данных – как задача выделения и распознавания новых (нештатных) состояний, то есть задача идентификации.

С помощью результатов решения всех вышеперечисленных задач можно осуществить и прогноз состояния системы.

Анализ ряда успешных применений Data Mining в технических приложениях позволяет утверждать, что методы Data Mining вполне применимы как средство анализа состояния «простой» системы и незаменимы при анализе «сложной» системы.

Однако в процессе поиска новых знаний в массиве информации применение самих методов Data Mining является лишь одним из его этапов, и они сами по себе не позволяют получить новые знания. Эффективное применение методов Data Mining предполагает наличие достаточно полных исходных данных для анализа. В случае анализа состояния сложных систем в качестве таковых выступают оценки контролируемых признаков, получаемые в результате обработки полученного объема измерительной информации, а также информации об условиях и режимах функционирования наблюдаемых систем, синхронизированной во времени. Однако при возникновении нештатных ситуаций возникает задача

оперативного получения оценок дополнительных контролируемых признаков из принятого объема измерительной информации для обеспечения полноты исходных данных для анализа.

Инструментарием, позволяющим организовать подготовку исходных данных, их оперативное восполнение и применение методов Data Mining, являются среды измерительного программирования, например LabVIEW [3]. Такая среда в рамках организации системы анализа позволяет решать следующие задачи:

- организация первичной измерительной системы без внесения технических изменений в объект контроля с целью получения исходной измерительной и сигнальной информации о его состоянии;
- формирование информационно-адресной системы регистрации данных с целью их дополнительной обработки методами Data Mining на момент нештатной ситуации;
- формирование базы знаний и данных исходных средств обработки измерительной и сигнальной информации с целью выбора и реализации гибких технологий интеллектуального анализа;
- организация системы всестороннего автоматизированного анализа исходной информации с целью выявления скрытых закономерностей на момент нештатной ситуации.

Перечисленные возможности сред измерительного программирования позволяют оперативно решать задачи получения оценок дополнительных контролируемых признаков состояния сложных систем в случае невыполнения требований по полноте исходной информации для проведения анализа средствами Data Mining.

В силу этого, перспективные аналитические платформы анализа состояния сложных технических систем должны быть совместимы с существующими компьютерными средствами измерений, реализующими новые измерительные технологии, например в рамках сред измерительного программирования.

Представляется, что в этом случае специалисты-анализаторы получат мощное средство поддержки и принятия решений, позволяющее обеспечить выполнение требований по оперативности и достоверности анализа сложных технических систем не только в штатных, но и при возникновении нештатных ситуаций.

Литература

1. Потюпкин А.Ю. Научно-методические основы решения задач анализа состояния объектов ракетно-космической техники в условиях неопределенности – М.: ВА РВСН, 2003.

2. Н. Паклин, В.Орешков. Бизнес-аналитика: от данных к знаниям. – СПб.: Питер, 2009.
3. http://www.labview.ru.

ИССЛЕДОВАНИЕ СОЦИАЛЬНО-ЭКОНОМИЧЕСКОГО РАЗВИТИЯ СЕЛЬСКИХ ТЕРРИТОРИЙ МЕТОДОМ КЛАСТЕРНОГО АНАЛИЗА

***Евтеева А.**, студентка*, ***Карпузова В.И.**, доцент**, Тарасова О.Б.**, профессор Российского государственного аграрного университета – МСХА имени К.А.Тимирязева, г. Москва*

Целью работы является разбиение сельских территорий Московской области (МО) на кластеры по комплексу существенных социально-экономических показателей и выявление, какие из них являются **наименее и наиболее** развитыми, а какие занимают срединные позиции.

Тема является актуальной, так как картина развития сельских территорий может быть передана наиболее точно лишь при рассмотрении всех факторов социально-экономического характера в совокупности, а объективная и полная информация – основа для принятия верных управленческих решений относительно устойчивого развития сельских территорий.

Для достижения поставленной в работе цели используется рабочее место аналитика *Deductor Studio*, которое входит в состав аналитической платформы *Deductor*. Данное приложение содержит набор механизмов импорта, обработки, визуализации и экспорта данных для быстрого и эффективного анализа информации. Для решения проблемы доклада была использована одна из возможностей аналитической платформы, а именно – карта Кохонена, которая позволяет проводить кластеризацию данных.

Особенностью кластерного анализа, как метода многомерной группировки является, то, что классифицируются многомерное наблюдения, каждое из которых описывается набором исходных переменных. Целью кластерного анализа является образование групп схожих между собой объектов, которые принято называть кластерами.

Поскольку сельские территории Московской области занимают большую часть от всей ее площади, были проведены исследования,

имеющие цель установить, как развиты по социально-экономическим показателям сельские территории районов МО.

При проведении анализа были взяты данные по социально-экономическому развитию сельских поселений районов МО. По сельскому хозяйству были взяты данные ВСХП о поголовье скота (которое было переведено в условное поголовье) и показатели по растениеводству. По социальному развитию были взяты 17 наиболее важных показателей, связанные с образованием, культурой, здравоохранением и другими отраслями социального развития. Все признаки представлены относительными величинами интенсивности. Часть этих данных были переведена на 1000 человек сельских поселений района, а остальная часть показателей на 100 га площади этих селений. Такой перевод является важным для объективной оценки территорий по уровню их развития и комплексной обеспеченности объектами социально-культурного быта.

Далее полученные данные были подготовлены для обработки в аналитической платформе *Deductor* и экспортированы в программу.

Сначала был проведен кластерный анализ районов по развитию сельского хозяйства. В качестве выходного признака было взято поголовье крупного рогатого скота (КРС), так как животноводство – основная отрасль области. Результатом кластеризации является самоорганизующаяся карта Кохонена и выходная таблица с номерами кластеров (табл. 1).

Таблица 1

Выходная таблица с номерами кластеров

Наименование района	№ кластера
Наро-Фоминский район	5
Одинцовский район	5
Подольский район	5
Раменский район	5
Коломенский район	4
Можайский район	4
Истринский район	4
Дмитровский район	4
Луховицкий район	4
Зарайский район	3
Сергиево-Посадский район	3
Шаховской район	3
Ступинский район	3
Серебряно-Прудский район	3
Чеховский район	3
Волоколамский район	3
Ленинский район	3

Наименование района	№ кластера
Серпуховский район	2
Рузский район	2
Каширский район	2
Клинский район	2
Воскресенский район	2
Талдомский район	2
Озёрский район	2
Щёлковский район	2
Ногинский район	1
Егорьевский район	1
Солнечногорский район	1
Шатурский район	1
Павлово-Посадский район	1
Лотошинский район	0
Пушкинский район	0
Мытищинский район	0
Красногорский район	0

Сельские территории были разбиты на шесть кластеров (рис. 1). Градация цветов от синего к красному позволяет определить положение районов. Синий цвет – отсталые районы, красный – передовые.

Лидерами в сельском хозяйстве оказались Нарофоминский, Одинцовский, Подольский и Раменский районы.

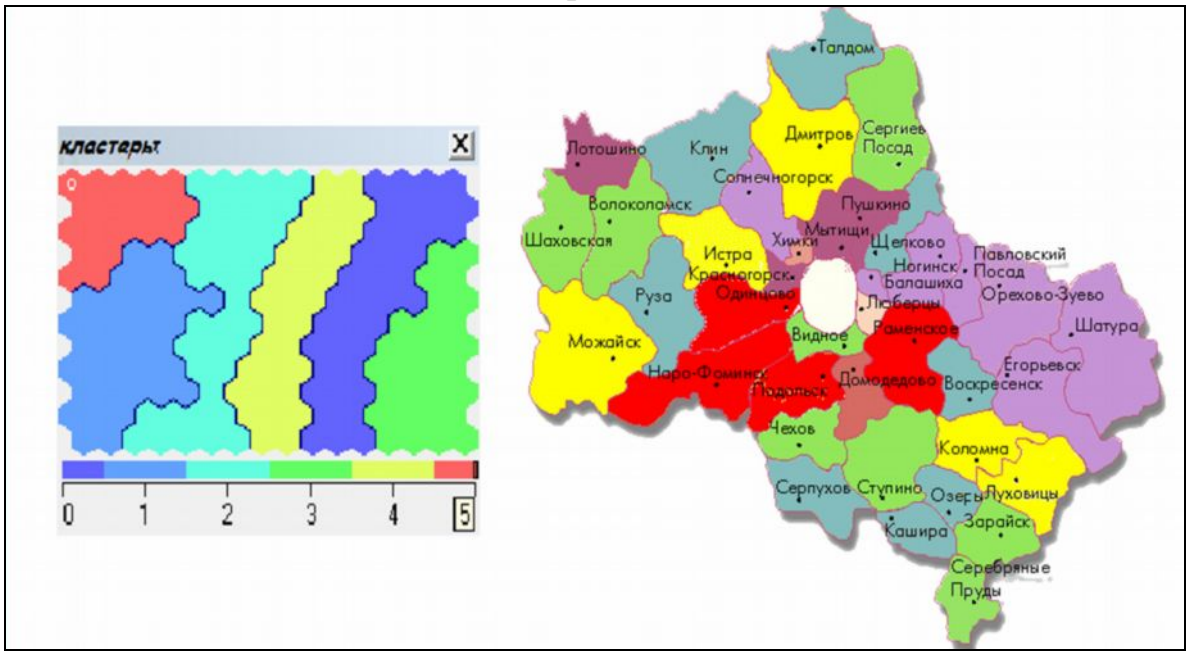


Рис. 1. Кластеризация районов Московской области по сельскохозяйственным показателям развития

Они попали в наивысший кластер. Аутсайдерами оказались Лотошинский, Пушкинский, Мытищинский и Красногорский районы. Как видно из рис. 1, наиболее передовые районы – это западные районы Московской области, так как в эти сельские поселения идет большая доля субсидий на сельское хозяйство.

Аналогичным образом с помощью аналитической *платформы Deductor* проведен кластерный анализ социально-бытового развития районов (табл. 2).

Результатом стало разбиение районов на шесть групп (рис. 2).

На карте (рис. 2) видно, что в самый развитый по всем показателям кластер попали районы, окружающие Москву. А самые неразвитые оказались на периферии области.

Таблица 2

Выходная таблица с номерами кластеров

Наименование района	№ кластера
Подольский район	5
Ленинский район	5
Одинцовский район	5
Истринский район	5
Раменский район	5
Дмитровский район	4
Солнечногорский район	4
Красногорский район	4
Шаховской район	4
Ногинский район	4
Клинский район	3
Наро-Фоминский район	3
Озёрский район	3
Коломенский район	3
Каширский район	3
Зарайский район	3
Щёлковский район	2
Серебряно-Прудский район	2
Луховицкий район	2
Чеховский район	2
Шатурский район	2
Воскресенский район	2
Мытищинский район	2
Сергиево-Посадский район	2
Пушкинский район	1
Можайский район	1
Волоколамский район	1
Серпуховский район	1
Ступинский район	1
Лотошинский район	0
Егорьевский район	0
Рузский район	0
Павлово-Посадский район	0

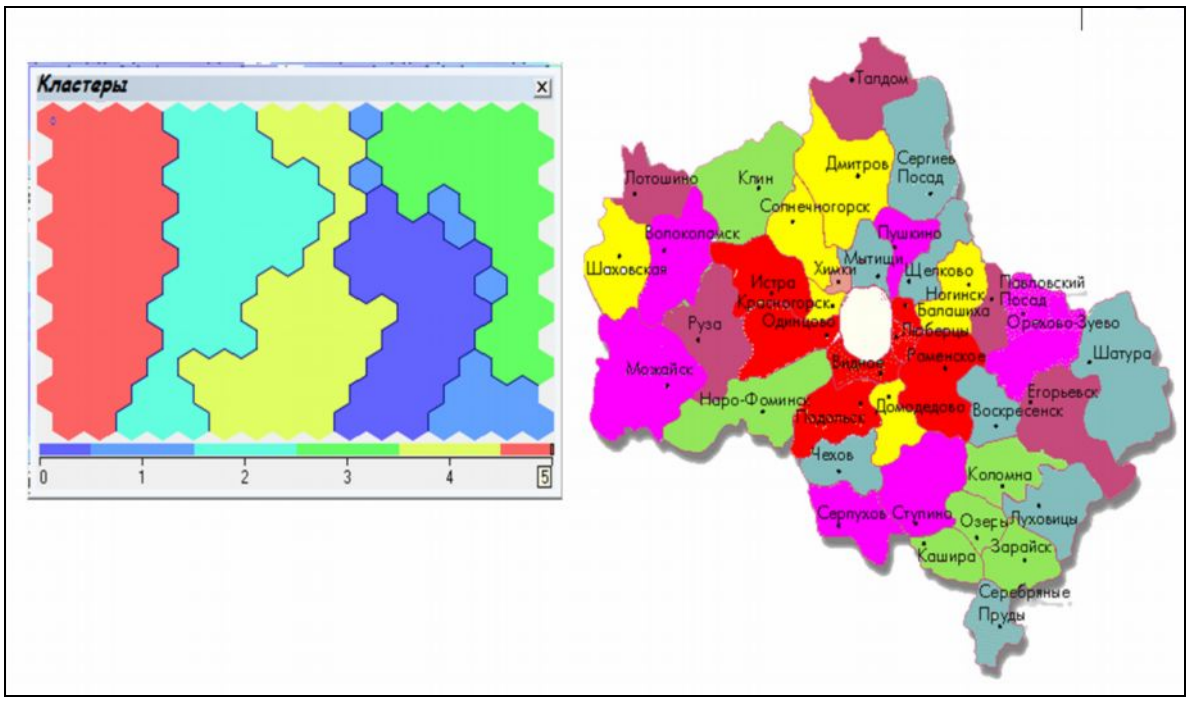


Рис. 2. Кластеризация районов Московской области по показателям социально-бытового развития

Далее было рассмотрено совокупное влияние сельскохозяйственных и социально-бытовых факторов на развитие сельских поселений. Для этого был проведен кластерный анализ по всем показателям (табл.

3). Результатом проведенного исследования были шесть кластеров сельских территорий (рис. 3).

Таблица 3

Выходная таблица с номерами кластеров

Наименование района	№ кластера
Одинцовский район	5
Подольский район	5
Раменский район	5
Солнечногорский район	4
Красногорский район	4
Шаховской район	4
Ленинский район	4
Дмитровский район	4
Коломенский район	4
Наро-Фоминский район	4
Клинский район	3
Истринский район	3
Ногинский район	3
Зарайский район	3
Луховицкий район	3
Чеховский район	2
Озёрский район	2
Серебряно-Прудский район	2
Каширский район	2
Шатурский район	2
Воскресенский район	2
Щёлковский район	2
Мытищинский район	2
Сергиево-Посадский район	2
Пушкинский район	1
Можайский район	1
Волоколамский район	1
Серпуховский район	1
Ступинский район	1
Лотошинский район	0
Егорьевский район	0
Рузский район	0
Павлово-Посадский район	0

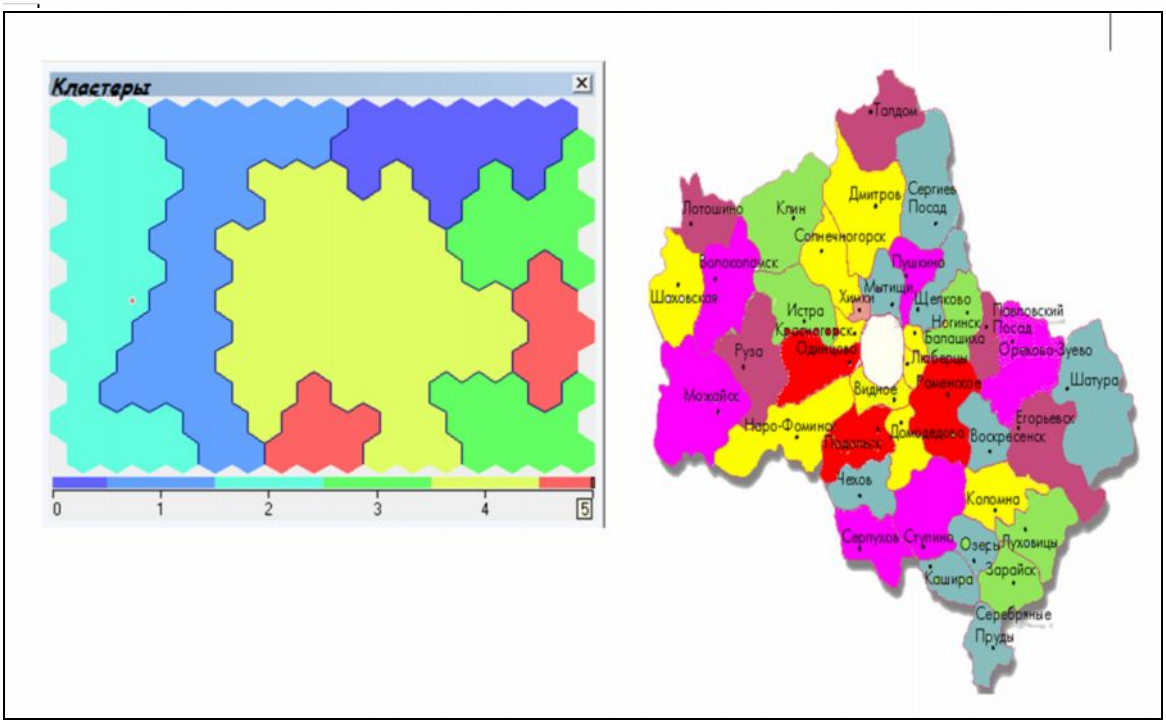


Рис. 3. Кластеризация районов Московской области по сельскохозяйственным и социально-бытовым показателям

Лидирующее положение имели три района, расположенные на юго-западе и в непосредственной близости к Москве. Это Одинцовский, Подольский, Раменский районы.

Четвертый кластер – имеет наибольший удельный вес районов в их общем количестве. Этот кластер характеризуется тем, что в нем высокие показатели и по сельскому хозяйству и довольно хорошие по социально-бытовому обеспечению.

Для последующих двух кластеров – третьего и второго характерны низкие показатели по развитию сельского хозяйства, но средние по социальному развитию.

Первый и нулевой кластер – аутсайдеры, эти районы находятся на окраинах Московской области. Этим районам необходима поддержка государства, как и в экономическую сторону, так и в социальную. В их отсталом от остальных районов развитии играет роль удаленность от Москвы.

Проведенное исследование позволило сделать следующие выводы:

- сельские поселения характеризуются системой трех взаимосвязанных групп показателей: развития сельского хозяйства, развития социально-культурной сферы жизни сельского населения, состояния экономики хозяйствующих субъектов;
- кластерный анализ по отдельным группам показателей позволяет выделить объективно существующие типы сельских поселений с позиций развития производственно-экономической деятельности на сельских территориях и социальной жизни сельского населения по всему комплексу признаков;
- типы поселений по производственным и социальным признакам указывают на тесную связь экономики и социально- бытовой инфраструктуры села;
- кластеризация сельских поселений по всему комплексу производственно-экономических и социально-культурных признаков обеспечила типизацию объекта исследования для принятия управленческих решений в вопросах развития сельских территорий.

ИСПОЛЬЗОВАНИЕ АНАЛИТИЧЕСКОЙ ПЛАТФОРМЫ DEDUCTOR ДЛЯ ИССЛЕДОВАНИЯ МИРОВОЙ КОНЪЮНКТУРЫ РЫНКА ПОДСОЛНЕЧНИКА

Мешков М., *студент*, ***Карпузова В.И.,*** *доцент****, Тарасова О.Б.,*** *профессор Российского государственного аграрного университета – МСХА имени К.А. Тимирязева, г. Москва*

В проведенном исследовании была поставлена цель – изучить мировой рынок подсолнечника, оценить по индикаторам конъюнктуры сложившийся тип рынка, выявить типы участников рынка подсолнечника и определить их динамическое развитие за период 1990-2008 гг., а также оценить роль России в формировании и развитии рынка.

Источниками информации послужили материалы ФАО. Основными методами при реализации цели были кластерный анализ стран участников рынка по комплексу показателей конъюнктуры рынка в сочетании с анализом рядов динамики на базе аналитической платформы *Deductor*.

В ходе работы для анализа мировой конъюнктуры по странам за период с 1990 по 2008 годы. изучены следующие показатели:

- посевные площади подсолнечника (га);
- урожайность (ц/га);
- цена за 1 тонну подсолнечника (USD/т);
- валовые сборы подсолнечника (тыс. т.);
- производство растительного масла (тыс. т.);
- экспорт семян подсолнечника (тыс. т.);
- импорт семян (тыс. т.).

Обобщение данные по странам мира показало, что количество участников рынка подсолнечника составляет 158 стран с небольшими колебаниями по годам.

Анализ индикаторов рынка показал, что рынок подсолнечника относится к типу развивающегося. Об этом свидетельствуют темпы роста экспортных и импортных операций, расширение посевов культур при росте урожайности культуры по большинству стран и, как следствие, расширение объемов производства продукции.

По посевным площадям базисный темп прироста в мире составил 46,8% (1990 – 2008 гг.), по урожайности – 6,7% (1990 – 2007 гг.), по производству – 57% (1990 – 2008 гг.), по экспорту 52,3% (1990 – 2008 гг.).

Среднегодовой прирост за те же промежутки времени по производству в мире составил 719 тыс. т., по экспорту 63 тыс. т., урожайности 0,1 ц/га.

В России прирост урожайности, также как и в мире, составил 0,1 ц/га, производства – 265 тыс. т., несмотря на сокращение экспорта (среднее абсолютное сокращение за рассматриваемый период –4,7 тыс. т.).

С ростом производства подсолнечника в мире увеличилось производство подсолнечного масла: базисный темп прироста составил 25,6% (1990 – 2008 гг.), среднегодовой абсолютный прирост – 163 тыс. т.

В России темп прироста составил – 61,7 %, что значительно выше мирового значения, среднегодовой абсолютный прирост – 83,3 тыс. т.

Дальнейший анализ показал, что на мировом рынке роль отдельных стран существенно различается. Основными производителями и экспортерами являются 15 стран, на долю которых приходится примерно 90 % всего производимого подсолнечника.

Учитывая наличие множества разнородных признаков-индикаторов рынка подсолнечника, был реализован метод кластеризации стран по основным параметрам. Кластеризация стран проводилась с использованием самоорганизующейся карты Кохонена. Кластеры на карте Кохонена разделены с учетом интенсивности изучаемых признаков: от худших по параметрам единиц совокупности до стран с самыми благоприятными параметрами развития рынка. В качестве основополагающего или выходного параметра был взят экспорт семян подсолнечника.

В результате кластеризации мы получили восемь кластеров стран с различными экономическими характеристиками (таблица 1).

Таблица 1

Сводная таблица по итогам кластеризации

Краткое описание кластера	**Количество стран в кластере**	
	1999	**2007**
Страны-импортеры	35	42
Страны-реэкспортеры (осуществляют экспортно-импортные операции без собственного производства)	19	29
Слабые участники мирового рынка подсолнечника (низкие цены производителя, минимум экспорта и импорта)	11	11
Слаборазвитый рынок, высокие цены	4	1
Крупные импортеры подсолнечника при собственном интенсивном производстве	23	26
Страны с интенсивным производством и минимальным импортом	23	24
Крупные производители с экстенсивным производством	–	3
Крупные производители с интенсивным производством и минимальным экспортом	5	–
Лидеры рынка: страны-экспортеры собственной продукции	3	5

Россия в 1999 г. входила в высшую группу благодаря большим размерам площадей под культурой, в 2007 г. перешла из высшего кластера в более низкий. Это связано, прежде всего, со снижением объемов экспорта и экстенсивными методами ведения сельского хозяйства.

Лишь большие площади посева культуры позволяют занимать России лидирующие позиции на мировом рынке.

Проведенное исследование позволяет сделать следующие выводы:

1. В настоящее время сложился развивающийся тип мирового рынка подсолнечника.

2. Россия находится в группе мировых лидеров, но из-за низкой интенсификации производства даже в годы относительно стабильного развития экономики России между кризисами 1998 и 2008 гг. теряла свои позиции на мировом рынке подсолнечника.

3. В России есть все предпосылки для успешного развития этой культуры (наличие природных условий, высокая обеспеченность рабочей силой в южных регионах страны, достаточно рынков сбыта). В этой связи необходимо рассматривать развитие производства подсолнечника в стране как экспортной культуры и интенсифицировать производство.

Исследование мировой конъюнктуры рынка подсолнечника при помощи кластерного анализа на базе аналитической платформы *Deductor* дает наиболее полную оценку происходящих на рынке динамических процессов, учитывая влияние всех факторов в совокупности.

АНАЛИТИЧЕСКАЯ ПЛАТФОРМА DEDUCTOR В ОЦЕНКЕ СТЕПЕНИ ЭКОЛОГИЧЕСКОЙ БЕЗОПАСНОСТИ РЕГИОНОВ

***Золотарева И.А.**, профессор, **Павленко Л.А.**, доцент, Харьковский национальный экономический университет, г. Харьков*

Сегодня общепризнано, что экономическое и социальное развитие, а также охрана окружающей природной среды, являются взаимосвязанными и взаимодополняющими компонентами устойчивого развития любого региона. Основным направлением программы ООН по экологической проблематике является система глобального экологического мониторинга, которая охватывает множество сетей наблюдения окружающей среды на международном и национальном уровнях [1]. Главными задачами экологического мониторинга являются: наблюдение за состоянием биосферы, оценка и прогноз ее состояния, определение степени антропо-

генного влияния на окружающую среду, выявление факторов и источников этого влияния, оптимизация отношений человека с природной средой, экологическая ориентация хозяйственной деятельности. Совершенствование антропогенного цикла связывают с двумя главными направлениями деятельности: мониторинг и управление (экономические мероприятия природоохранного регулирования) [2].

Решение этих задач невозможно без применения современных информационных технологий, позволяющих собирать, хранить, обрабатывать информацию о состоянии окружающей среды, в том числе топологическую информацию про размещение территориально распределенных объектов, принимать оперативные управленческие решения, распространять результаты анализа. Задачи, связанные с анализом данных, привязанных к картооснове, успешно решаются в рамках геоинформационных систем [3].

На кафедре информационных систем Харьковского национального экономического университета, в рамках дисциплины «Системы обработки эколого-экономической информации» со студентами специальности «Компьютерный эколого-экономический мониторинг» задачи анализа результатов мониторинга решаются с применением *Deductor Academic* [3] и пакета *ArcGis* [4].

В данной работе приводится пример использования аналитической платформы Deductor при решении задачи оценки степени экологической безопасности регионов Украины по данным Департамента региональной политики Министерства экономики по вопросам европейской интеграции Украины, размещаемым в INTERNET. Результаты разбиения множества данных на подмножества кластеров выведены на карте Украины.

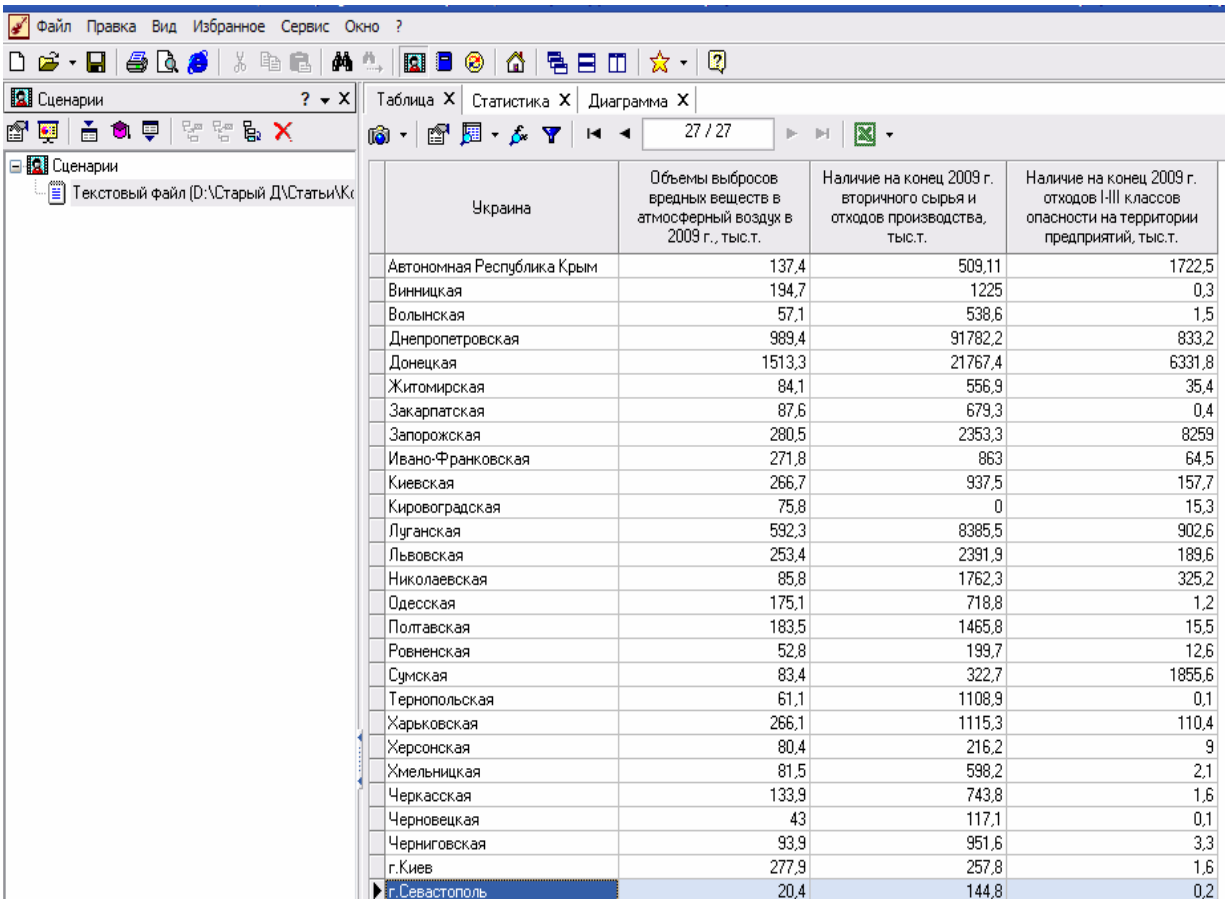
Согласно данным мониторинга каждый из регионов характеризуется совокупностью экологических, экономических, социальных показателей. В частности, известны данные о загрязнении регионов Украины на конец 2009 года, приведенные в табл. 1 (формат *.xls).

Таблица 1

Экологические показатели регионов Украины в 2009 году

Украина	**Объемы выбросов вредных веществ в атмосферный воздух в 2009 г., тыс.т.**	**Наличие на конец 2009 г. вторичного сырья и отходов производства, тыс.т.**	**Наличие на конец 2009 г. отходов I-III классов опасности на территории предприятий, тыс.т.**
Автономная Республика Крым	137,4	509,11	1722,5
Винницкая	194,7	1225	0,3
Волынская	57,1	538,6	1,5
Днепропетровская	989,4	91782,2	833,2
Донецкая	1513,3	21767,4	6331,8
Житомирская	84,1	556,9	35,4
Закарпатская	87,6	679,3	0,4
Запорожская	280,5	2353,3	8259
Ивано-Франковская	271,8	863	64,5
Киевская	266,7	937,5	157,7
Кировоградская	75,8	0	15,3
Луганская	592,3	8385,5	902,6
Львовская	253,4	2391,9	189,6
Николаевская	85,8	1762,3	325,2
Одесская	175,1	718,8	1,2
Полтавская	183,5	1465,8	15,5
Ровненская	52,8	199,7	12,6
Сумская	83,4	322,7	1855,6
Тернопольская	61,1	1108,9	0,1
Харьковская	266,1	1115,3	110,4
Херсонская	80,4	216,2	9
Хмельницкая	81,5	598,2	2,1
Черкасская	133,9	743,8	1,6
Черновецкая	43	117,1	0,1
Черниговская	93,9	951,6	3,3
г.Киев	277,9	257,8	1,6
г.Севастополь	20,4	144,8	0,2

Для разбиения исходного набора данных из табл. 1 на подмножества с целью отнесения каждого из регионов к одному из классов экологической безопасности был выполнен кластерный анализ. Для этого таблица была сохранена в формате *.txt и импортирована в аналитическую платформу *Deductor*. На рис. 1 представлено окно *Deductor Studio* с импортированными данными в табл. 1.



Украина	Объемы выбросов вредных веществ в атмосферный воздух в 2009 г., тыс.т.	Наличие на конец 2009 г. вторичного сырья и отходов производства, тыс.т.	Наличие на конец 2009 г. отходов I-III классов опасности на территории предприятий, тыс.т.
Автономная Республика Крым	137,4	509,11	1722,5
Винницкая	194,7	1225	0,3
Волынская	57,1	538,6	1,5
Днепропетровская	989,4	91782,2	833,2
Донецкая	1513,3	21767,4	6331,8
Житомирская	84,1	556,9	35,4
Закарпатская	87,6	679,3	0,4
Запорожская	280,5	2353,3	8259
Ивано-Франковская	271,8	863	64,5
Киевская	266,7	937,5	157,7
Кировоградская	75,8	0	15,3
Луганская	592,3	8385,5	902,6
Львовская	253,4	2391,9	189,6
Николаевская	85,8	1762,3	325,2
Одесская	175,1	718,8	1,2
Полтавская	183,5	1465,8	15,5
Ровненская	52,8	199,7	12,6
Сумская	83,4	322,7	1855,6
Тернопольская	61,1	1108,9	0,1
Харьковская	266,1	1115,3	110,4
Херсонская	80,4	216,2	9
Хмельницкая	81,5	598,2	2,1
Черкасская	133,9	743,8	1,6
Черновецкая	43	117,1	0,1
Черниговская	93,9	951,6	3,3
г.Киев	277,9	257,8	1,6
г.Севастополь	20,4	144,8	0,2

Рис. 1. Результат импорта данных табл. 1 в Deductor

На рис. 2 представлено окно настройки назначения столбцов для выполнения кластерного анализа.

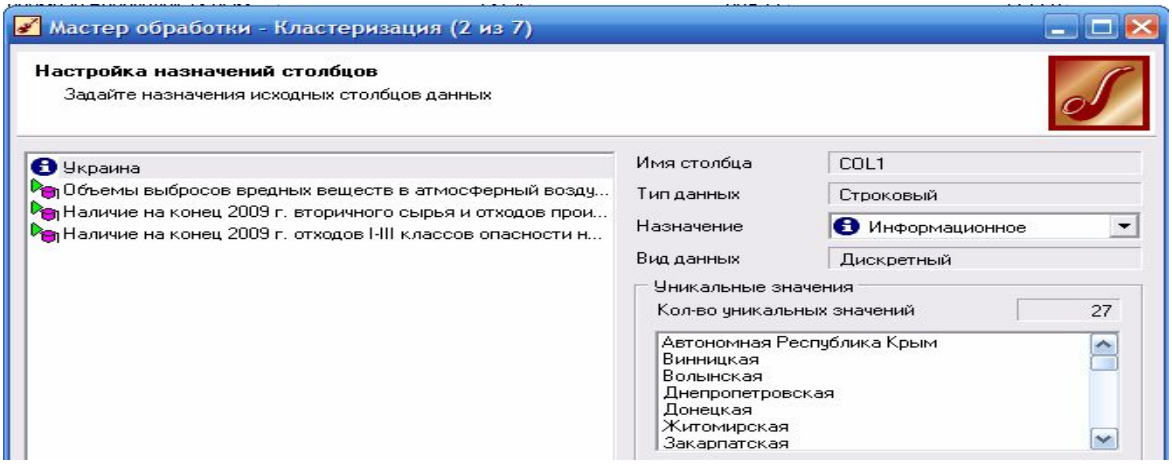


Рис. 2. Окно настройки назначения столбцов

Для обеспечения наглядности представления полученных результатов была выбрана кубическая модель данных. На рис. 3 представлено окно настройки полей куба.

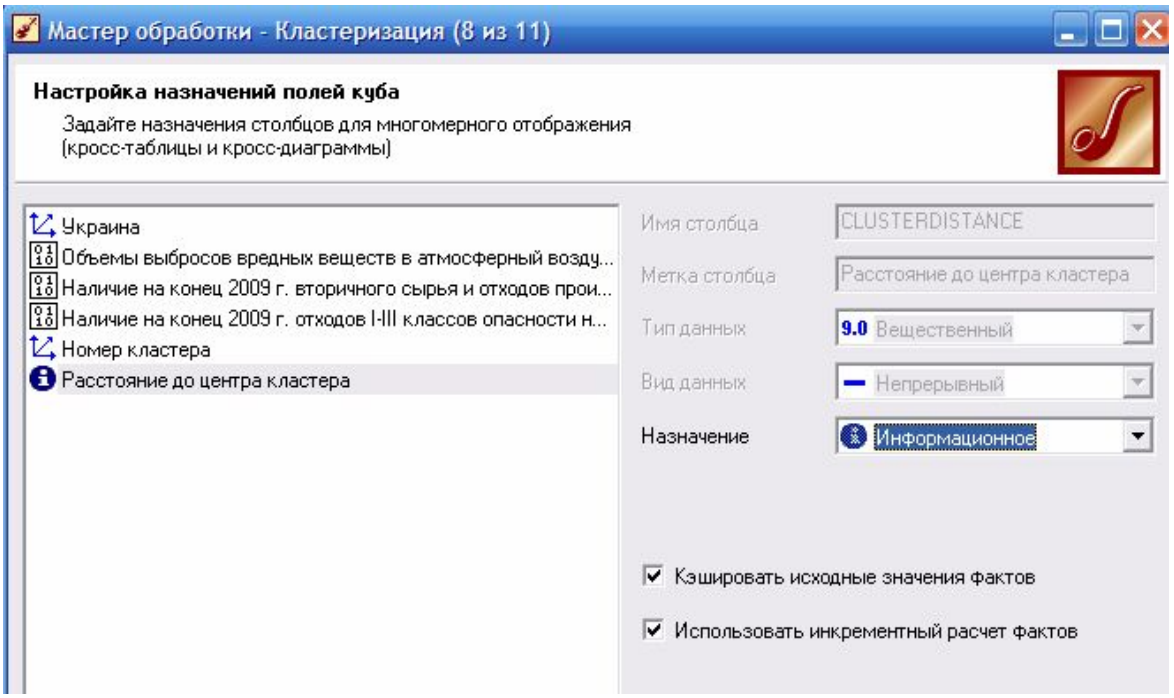


Рис. 3. Окно настройки полей куба

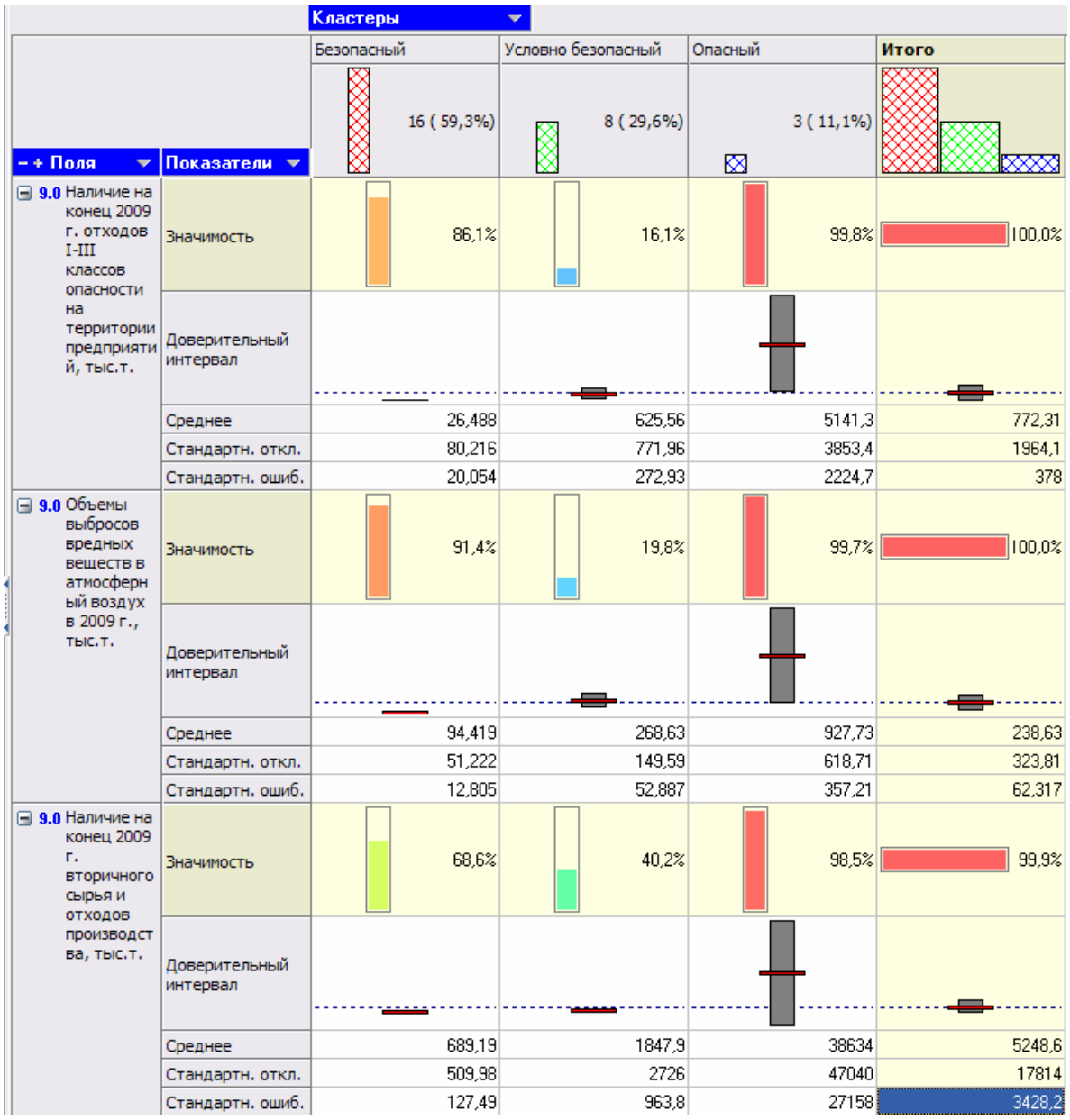
Пакет позволяет выбрать один из методов кластеризации: *g-means* или *k-means* и дать наименования полученным подмножествам данных. В данном случае был выбран метод *k-means* с разбиением на три кластера.

На рис. 4 представлено окно результатов кластеризации – профили кластеров с достигнутой точностью решения.

Группы регионов в экологическом отношении были названы: кластер 0 – «Безопасный», кластер 1 – «Условно безопасный», кластер 2 – «Опасный».

На рис. 5 приведен нормализованный график показателей экологической безопасности регионов Украины по трем показателям: «Объемы выбросов вредных веществ в атмосферный воздух», «Наличие на конец 2009 г. вторичного сырья и отходов производства», «Наличие на конец 2009 г. отходов I-III классов опасности на территории предприятий».

Исходные данные на графике приведены к единому масштабу для наглядности. Пакет позволяет представить данные, как в естественном виде, так и нормализованными, что повышает оперативность анализа.



Кластеры		Безопасный	Условно безопасный	Опасный	Итого
		16 (59,3%)	8 (29,6%)	3 (11,1%)	
Поля	Показатели				
9.0 Наличие на конец 2009 г. отходов I-III классов опасности на территории предприятий, тыс.т.	Значимость	86,1%	16,1%	99,8%	100,0%
	Доверительный интервал				
	Среднее	26,488	625,56	5141,3	772,31
	Стандартн. откл.	80,216	771,96	3853,4	1964,1
	Стандартн. ошиб.	20,054	272,93	2224,7	378
9.0 Объемы выбросов вредных веществ в атмосферный воздух в 2009 г., тыс.т.	Значимость	91,4%	19,8%	99,7%	100,0%
	Доверительный интервал				
	Среднее	94,419	268,63	927,73	238,63
	Стандартн. откл.	51,222	149,59	618,71	323,81
	Стандартн. ошиб.	12,805	52,887	357,21	62,317
9.0 Наличие на конец 2009 г. вторичного сырья и отходов производства, тыс.т.	Значимость	68,6%	40,2%	98,5%	99,9%
	Доверительный интервал				
	Среднее	689,19	1847,9	38634	5248,6
	Стандартн. откл.	509,98	2726	47040	17814
	Стандартн. ошиб.	127,49	963,8	27158	3428,2

Рис. 4. Результат кластеризации регионов Украины по экологическим показателям

На рис. 6 представлен результат построения куба с данными разбиения регионов Украины по трем показателям, характеризующим экологическое состояние регионов.

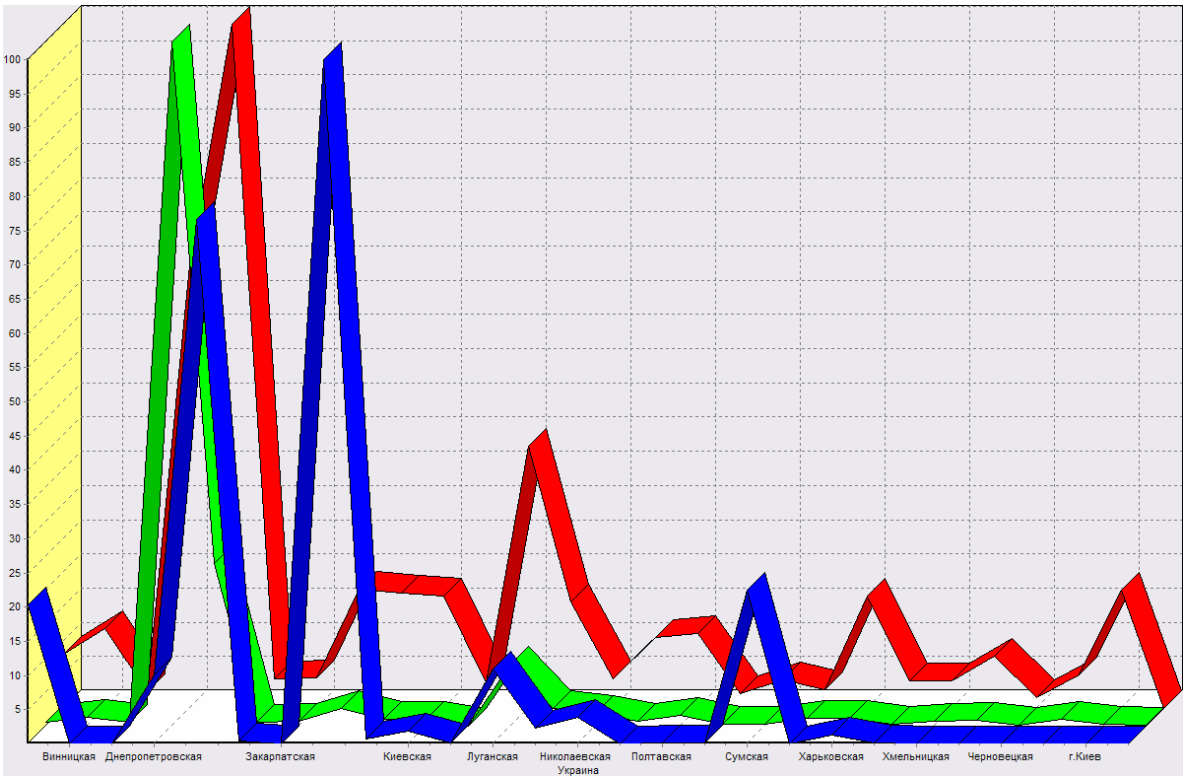


Рис. 5. Диаграмма экологических показателей регионов Украины с нормализованными данными

Модель рис. 6 наглядно представляет экологические показатели каждой из групп, позволяет суммировать их по группе и принимать решения, как по отдельным показателям в регионах, так и по всем показателям каждого региона.

Дополнением решения этой задачи является вывод результатов кластеризации на карте Украины. Для этого результаты кластеризации – данные таблицы рис. 4 были сохранены в пакете *Excel*. Средствами Excel данные этой таблицы были сохранены в формате *.dbf и соединены с таблицей регионов пакета *ArcViewGIS 3.3*. На рис. 7 представлена карта Украины, построенная в среде *ArcViewGIS 3.3* с результатами оценки экологической безопасности регионов.

В данном случае это просто отражение факта отнесения регионов к определенному классу экологической безопасности. На самом деле аналитические возможности пакета *ArcGIS* позволяют его использовать для решения многих задач анализа и принятия управленческих решений.

	Номер кластера								
	0			1			2		
Украина	Σ Объемы вы	Σ Наличие на кс	Σ Наличие на кои	Σ Объемы вы	Σ Наличие на кс	Σ Наличие на кои	Σ Объемы вы	Σ Наличие на кс	Σ Наличие на кои
Автономная Республика Крым				137,40	509,11	1 722,50			
Винницкая	194,70	1 225,00	0,30						
Волынская	57,10	538,60	1,50						
Днепропетровская							989,40	91 782,20	833,20
Донецкая							1 513,30	21 767,40	6 331,80
Житомирская	84,10	556,90	35,40						
Закарпатская	87,60	679,30	0,40						
Запорожская							280,50	2 353,30	8 259,00
Ивано-Франковская				271,80	863,00	64,50			
Киевская				266,70	937,50	157,70			
Кировоградская	75,80	0,00	15,30						
Луганская				592,30	8 385,50	902,60			
Львовская				253,40	2 391,90	189,60			
Николаевская	85,80	1 762,30	325,20						
Одесская	175,10	718,80	1,20						
Полтавская	183,50	1 465,80	15,50						
Ровненская	52,80	199,70	12,60						
Сумская				83,40	322,70	1 855,60			
Тернопольская	61,10	1 108,90	0,10						
Харьковская				266,10	1 115,30	110,40			
Херсонская	80,40	216,20	9,00						
Хмельницкая	81,50	598,20	2,10						
Черкасская	133,90	743,80	1,60						
Черниговская	93,90	951,60	3,30						
Черновецкая	43,00	117,10	0,10						
г.Киев				277,90	257,80	1,60			
г.Севастополь	20,40	144,80	0,20						
Итого:	1 510,70	11 027,00	423,80	2 149,00	14 782,81	5 004,50	2 783,20	115 902,90	15 424,00

Рис. 6. Кубическая модель данных с результатами разбиения регионов Украины на кластеры по экологическим показателям

Рис. 7. Карта Украины с результатами оценки экологической безопасности регионов

Аналитическая платформа *Deductor* позволяет быстро, качественно, в комфортной для конечного пользователя форме выполнить кластерный анализ и представить данные в виде многомерной модели, позволяющей оперативно анализировать и принимать управленческие решения, в данном случае, с целью улучшения экологического состояния регионов и государства в целом. Достоинством платформы также является возможность выполнения обмена данными и результатами анализа со многими инструментальными средствами обработки данных других фирм.

Литература

1. Богобоящий В.В., Чурбанов К.Р., Палій П.Б., Шмандій В.М. Принципи моделювання та прогнозування в екології: Підручник. — Київ: Центр навчальної літератури, 2004. – 216 с.

2. Марчук Г.И. Математическое моделирование в проблеме окружающей среды. — М.: Наука, 1982. – 320 с.

3. http://www.basegroup.ru/download/deductor/.

4. http://gis.report.ru.

ИСПОЛЬЗОВАНИЕ НЕЙРОННЫХ СЕТЕЙ ДЛЯ АНАЛИЗА ЭКОНОМИЧЕСКИХ ПОКАЗАТЕЛЕЙ РЕГИОНОВ ПРИВОЛЖСКОГО ФЕДЕРАЛЬНОГО ОКРУГА

***Шамсутдинова Т.М.**, доцент Башкирского государственного аграрного университета,*

***Мухаметшин Т.Р.**, студент Башкирского государственного аграрного университета, г. Уфа*

Целью исследования являлось построение имитационной модели в виде нейронной сети, позволяющей провести анализ основных социально-экономических показателей регионов Приволжского федерального округа.

При проведении исследования были использованы следующие виды статистических данных за январь 2010 года по 14 областям, краям и республикам данного федерального округа[1]:

- объем выполненных работ по виду деятельности «Строительство», млн.руб.;

[1] По материалам статистических отчетов http://www.bashstat.ru/bashdigital/region16/default.aspx

- ввод в действие жилых домов, кв. м общей площади;
- оборот розничной торговли, млн. руб.;
- объем платных услуг населению с учетом неформального сектора экономики, млн. руб.;
- общий объем оборота оптовой торговли, млн. руб.;
- среднемесячная заработная плата работников, руб.;
- сводный индекс потребительских цен на товары и платные услуги (январь 2010 г. в % к декабрю 2009 г.);
- индекс цен производителей промышленных товаров (январь 2010 г. в % к декабрю 2009 г.);
- рост заработной платы (январь 2010 г. в % к январю 2009 г.);
- индекс производства по видам деятельности «Добыча полезных ископаемых», «Обрабатывающие производства» и «Производство и распределение электроэнергии, газа и воды» (январь 2010 г. в % к январю 2009 г.).

Так как регионы Приволжского федерального округа достаточно сильно отличаются размерами своих территорий, то для получения более корректных результатов анализа данных при проведении исследования также были использованы сведения о площадях территорий регионов[1] данного округа (тыс. кв. км).

Для построения нейронной сети был использован узел *Нейросеть* аналитической платформы *Deductor*.

При обучении нейронной сети были использованы результаты кластерного анализа данных в виде самоорганизующихся карт Кохонена. Для построения и обучения самоорганизующихся карт использовался инструмент анализа *Карта Кохонена* аналитической платформы *Deductor*.

С использованием построенных карт Кохонена был проведен кластерный анализ данных, в результате которого были выявлены группы регионов с различными и сходными социально-экономическими показателями.

При этом в кластер № 0, расположенный в правой зоне самоорганизующихся карт (рис. 1), вошли Пермский край, Кировская область, Республика Марий Эл и Республика Мордовия. Данный кластер характеризуется высоким индексом роста производства по видам деятельности «Добыча полезных ископаемых», «Обрабатывающие производства» и «Производство и распределение электроэнергии, газа и воды», но, при этом, достаточно низким объемом выполненных работ по виду деятель-

[1] По материалам http://worldgeo.ru/russia/okr_privolg/, раздел «Площади территорий регионов Приволжского федерального округа»

ности «Строительство», сравнительно небольшим оборотом розничной торговли и высоким индексом роста цен производителей промышленных товаров.

Из регионов, вошедших в данный кластер, наилучшие показатели по большинству позиций имеет Пермский край, причем этот край – один лидеров по уровню среднемесячной заработной платы во всем Приволжском округе. Но надо отметить, что Пермский край является самым крупным регионом округа (около 160 тыс. кв. км), в то время как Республика Марий Эл и Республика Мордовия имеют территорию всего 23 тыс. кв. км и 26 тыс. кв. км соответственно.

К кластеру № 1, расположенному в центральной зоне карт Кохонена, были отнесены Нижегородская, Оренбургская, Самарская и Саратовская области, Республика Башкортостан, Республика Татарстан и Чувашская Республика. Большинство регионов данного кластера занимают лидирующее место по таким показателям, как объем выполненных работ по виду деятельности «Строительство», ввод в действие жилых домов, оборот розничной торговли, объем платных услуг населению с учетом неформального сектора экономики, общий объем оборота оптовой торговли, рост заработной платы. Но при этом у данных регионов достаточно высокий сводный индекс роста потребительских цен на товары и платные услуги и высокий индекс роста цен производителей промышленных товаров.

Среди регионов данного кластера наилучшие экономические показатели имеют Республика Татарстан, Республика Башкортостан, Нижегородская и Самарская области. Наихудшие показатели по объему выполненных работ по виду деятельности «Строительство», по обороту розничной и оптовой торговли – у Республики Чувашия. Но при этом следует учесть, что данная республика имеет самую маленькую территорию (по сравнению с другими регионами округа) – всего 18 тыс. кв. км.

К кластеру № 2, находящемуся в левой зоне карт Кохонена, были отнесены Удмуртская Республика, Пензенская и Ульяновская области. Все эти регионы имеют небольшую территорию – от 37 до 43 тыс. кв. км. Для них характерны средние значения показателей по вводу в действие жилых домов, по обороту розничной торговли, низкий индекс роста производства по видам деятельности «Добыча полезных ископаемых», «Обрабатывающие производства» и «Производство и распределение электроэнергии, газа и воды». При этом в данных регионах средний для Приволжского федерального округа уровень заработной платы, но достаточно высокий сводный индекс роста потребительских цен на товары и платные услуги.

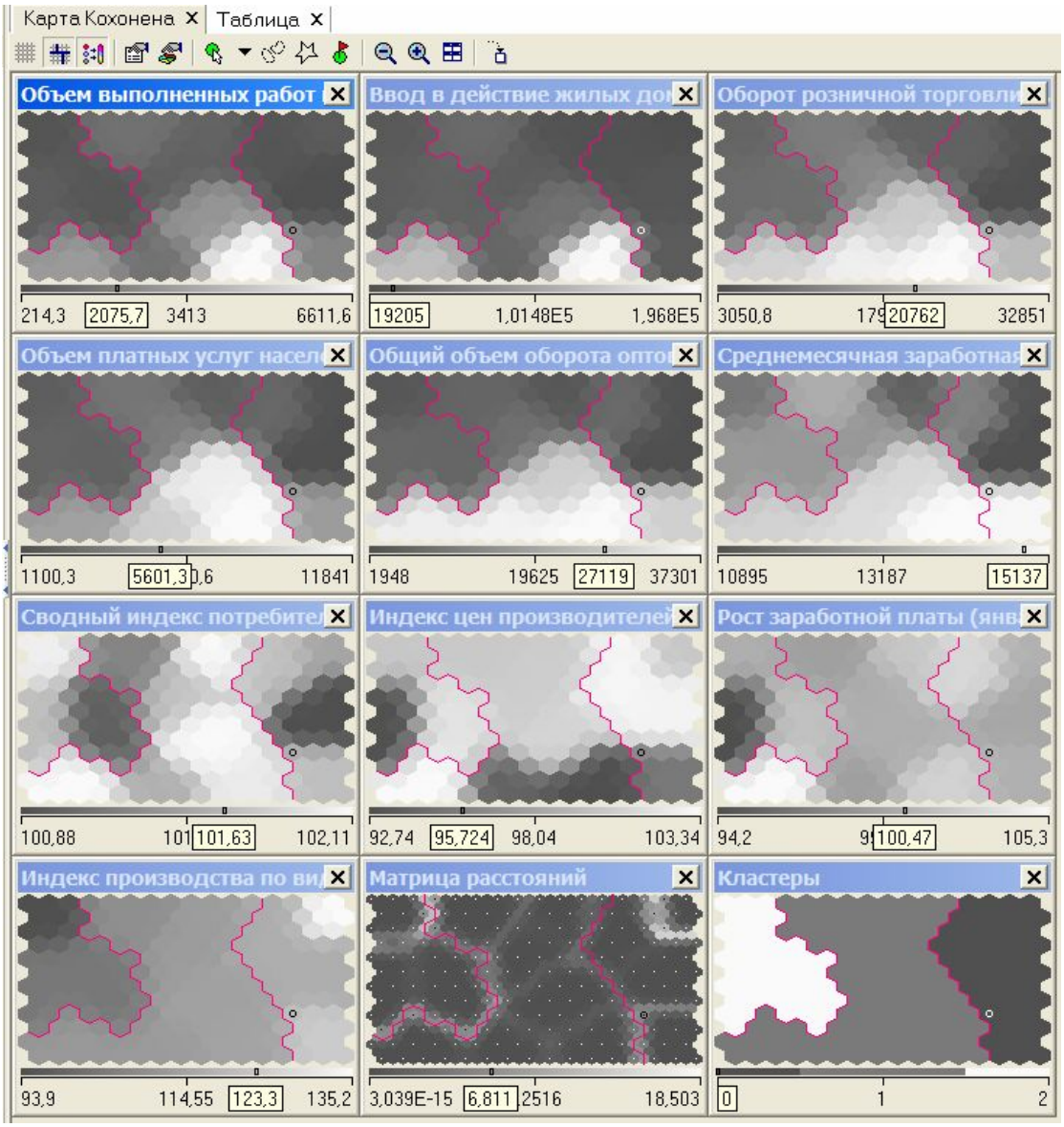


Рис. 1. Результаты кластерного анализа данных

После проведения анализа карт Кохонена была построена нейронная сеть, позволяющая проанализировать влияние отдельных факторов на отнесение региона к тому или иному кластеру.

Настройка и обучение нейронной сети в Deductor Studio состоит из следующих шагов: настройка назначений полей и их нормализация, настройка обучающей выборки, настройка структуры нейронной сети, выбор алгоритма и параметров обучения, настройка условий остановки обучения, запуск процесса обучения, выбор способа отображения данных. В качестве способов отображения нейронной сети в данном случае были выбраны «Граф нейросети» и визуализатор «Что-если» (рис. 2).

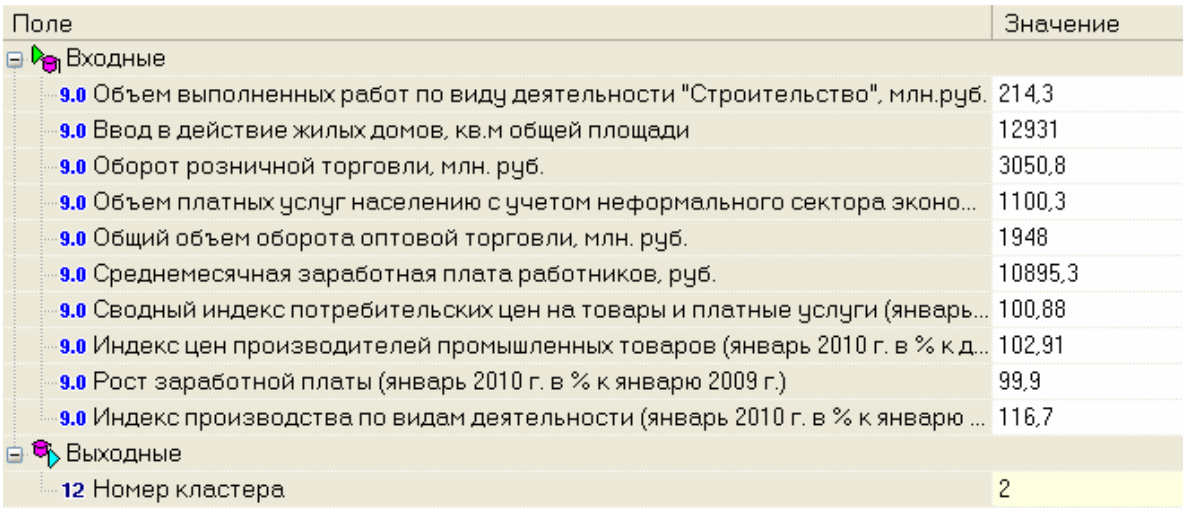


Рис. 2. Визуализатор нейросети «Что-если»

Интерактивные эксперименты в визуализаторе «Что-если» позволил проанализировать влияние отдельных факторов на отнесение региона к тому или иному кластеру.

В результате анализа данных было получено, что наиболее значимыми факторами для принятия решения об отнесении региона к определенному кластеру являются такие критерии как:

1. объем выполненных работ по виду деятельности «Строительство», млн.руб.;
2. индекс производства по видам деятельности «Добыча полезных ископаемых», «Обрабатывающие производства» и «Производство и распределение электроэнергии, газа и воды» (январь 2010 года в % к январю 2009 года).

По первому критерию лучшие показатели среди регионов Приволжского федерального округа были выявлены у республик и областей, относящихся к кластеру №1, по второму критерию – у регионов, относящихся к кластеру №0. В кластер № 2 вошли регионы, имеющие средние показатели по первому критерию, но отстающие по второму критерию.

Итак, были реализованы следующие этапы анализа данных:

- на основании открытой статистической отчетности был сформирован набор данных, содержащий ряд основных социально-экономических показателей по 14 областям, краям и республикам Приволжского федерального округа;
- с использованием инструмента анализа «Карта Кохонена» *Deductor Studio* был проведен кластерный анализ данных, в результате которого были выявлены группы регионов с различным уровнем социально-экономических показателей;

- с использованием инструмента анализа «Нейросеть» *Deductor* Studio была построена нейронная сеть, позволяющая проанализировать влияние отдельных факторов на отнесение региона к тому или иному кластеру. При обучении нейронной сети были использованы результаты кластерного анализа данных в виде самоорганизующихся карт Кохонена. В результате анализа данных были выявлены наиболее значимые факторы для принятия решения об отнесении региона к определенному кластеру.

Полученная методика может быть применена при анализе социально-экономических показателей Приволжского федерального округа, а также может быть использована специалистами в процессе принятия управленческих решений при проведении экономической политики в регионах данного округа. Разбиение на кластеры позволит выявить слабые показатели регионов и принять решение об изменении стратегии их развития с целью улучшения их социально-экономического положения.

Данная методика может быть применена и для анализа социально-экономических показателей других федеральных округов, а также для анализа результатов деятельности экономических субъектов в регионах.

СЕГМЕНТАЦИЯ КЛИЕНТОВ БРОКЕРСКОГО ОБСЛУЖИВАНИЯ

***Нейский И.М.**, аспирант, **Филиппович А.Ю.**, доцент, Московский государственный технический университет им. Н.Э. Баумана*

Большинство современных предприятий используют в своей деятельности информационные системы, хранилища данных, в которых собираются данные по бизнес – процессам компании. Объемы накапливаемой информации увеличиваются с течением времени, поэтому актуальной задачей в развитии компании является переход от анализа тенденций текущих показателей деятельности предприятия к более комплексном подходу «извлечения знаний» из имеющихся данных в целях выявления закономерностей.

Изучением проблем и созданием решений в этой области активно занимаются направления Интеллектуального анализа данных (Business Intelligence) и Управления знаниями (Knowledge Management), в рамках которых выделяются поднаправления Выявление знаний в базах данных (Knowledge Discovery in Databases), Анализ фактографических данных

(Data Mining), Анализ неструктурированных данных (Text Mining) и др. Результаты исследований этих направлений положены в основу многих информационно-аналитических систем, которые используются, в основном, для персональной работы экспертов. Однако современной тенденцией является применение указанных технологий и для централизованного управления организациями.

Для исследования структурированных массивов информации используется анализ фактографических данных, в котором выделены шесть различных задач: классификация, регрессия, кластеризация, выявление ассоциаций, выявление последовательностей, и прогнозирование. Потребность в кластеризации возникает в тех областях/этапах деятельности, где есть необходимость в разбиении объектов (ситуаций) на непересекающиеся подмножества, называемыми кластерами, так, чтобы каждый кластер состоял из схожих объектов, а объекты разных кластеров существенно отличались. Четкое разделение на кластеры возможно только в идеальных условиях и при сильно различающихся параметрах объектов кластеризации, поэтому для решения реальных задач все чаще применяются нечеткие методы, в которых разбиение объектов (ситуаций) выполняется на частично пересекающиеся подмножества.

Примером организаций, для которых актуальна проблема анализа накопленной информации, являются финансовые компании – профессиональные участники, работающие на фондовом рынке, которые привлекают клиентов на брокерское обслуживание. На сегодняшний день в России существует более 60 крупных компаний со среднемесячным оборотом около 800 миллионов долларов США [1]. Основным показателем эффективности работы в данном направлении является объем клиентских оборотов и комиссионных сборов за совершаемые от их имени и за их счет операции, поэтому для успешного развития брокерского обслуживания необходимо увеличивать количество клиентов и/или их обороты, на основе которых, как правило, определяется сумма комиссионного вознаграждения. Ввиду того, что каждый клиент по-своему уникален (желания, возможности, предпочтения, стратегия и т.п.), то для его привлечения и создания заинтересованности от компании требуется существенная гибкость. Так как отвечать интересам каждого клиента со стороны крупной компании, обслуживающей более 10 000 клиентов, практически невозможно при текущем штате сотрудников компании, которые сопровождают заключение и оперативную обработку этих операций, поэтому для обеспечения дальнейшего развития компании проводится анализ клиентской базы с целью выделения характерных групп клиентов. По результатам данного исследования для полученных групп клиентов разрабатываются индивидуальные тарифы, условия обслуживания и т.д.

Данный подход подтверждает свою эффективность даже в кризисный период, так как при падении на мировых финансовых рынках экономических показателей происходит отток капитала и клиентов. Создание индивидуальной продуктовой линейки и снижение необходимости значительного расширения штата сотрудников для сопровождения новых операций клиентов ведет к снижению издержек, а значит и к снижению тарифов при их обслуживании, что привлекает новых и расширяет количество операций существующих клиентов. Кризис – это не только спад показателей в различных отраслях экономики, но и возможность достичь более значимых результатов за счет повышения собственной эффективности, поэтому появляется необходимость использовать дополнительные, ранее не используемые ресурсы, которые сосредоточены в компании – внутреннее информационное поле (аналитики, эксперты, накопленная информация об операциях, клиентах и т.д.).

Главная особенность анализа этой области в том, что его необходимо проводить на регулярной основе, чтобы сохранить конкурентные преимущества на рынке данного вида услуг. Учитывая динамику роста клиентской базы, использование ранее применяемых методов с привлечением только человеческих ресурсов становится невозможно, так как объем информации для анализа также возрастает. На данный момент известно более 100 методов кластеризации, поэтому для перехода на использование машинных методов необходимо осуществить выбор метода или методов из существующих либо разработать собственный метод с учетом особенностей этой области. Анализ существующих решений и методов [2] показал, что на текущий момент не существует специализированных или успешно примененных универсальных методов для решения описанной задачи. Еще одной проблемой в данной области является оценка качества получаемого результата и выбор количества групп – кластеров, которое является входным параметром для большинства алгоритмов.

В связи с тем, что на данный момент не существует достаточного количества практических рекомендаций по применению существующих методов в данной предметной области и количество методов достаточно велико, была разработана методика адаптивной кластеризации, которая направлена на решение этой задачи. Данная методика, состоящая из четырех этапов, позволяет осуществить выбор метода кластерного анализа и получить конечное разбиение множества исходных объектов на кластеры. На основе методики получено, что для решения задачи разбиения клиентов брокерского обслуживания необходимо разработать новый метод адаптивной кластеризации, в котором количество кластеров является результатом исследования.

После проведенного анализа для решения поставленной задачи из инструментов выполнения кластеризации были выбраны: теория графов и нечеткая логика. Определяющими факторами в выбранной комбинации является способность при использовании графов выделять кластеры произвольной формы и оптимальной структуры, а при использовании математического аппарата нечеткой логики решается задача разделения объектов с лингвистическими атрибутами. За основу для нового метода в части первичного разделения объектов на кластеры взята идея метода MST [3], использующего минимальные остовные деревья, и идея метода Fuzzy C-means [4]. На базе этих методов разработан метод ADAKL (рис. 1), который является двухэтапным и использует оценочную функцию разбиения, повышающую качество проводимой кластеризации [5]. Вычисление глобального критерия делает алгоритм кластеризации во много раз быстрее, чем при использовании локального критерия при парном сравнении объектов.

При работе ADAKL строится минимальное остовное дерево, образуя оптимизированную древовидную структуру из исходных элементов на основе характеристик кластеризуемых объектов, и выделяются первичные кластерные центры. Затем используется итерационный подход, с помощью которого уточняются центры кластеров и содержимое кластеров на основе вычисления степени принадлежности объекта кластеру.

Для устранения чувствительности к выбросам на первом этапе предлагается использование предобработки исходных данных через линейную или статистическую нормализацию. При вычислении информационных расстояний между объектами используются классические метрики, доработанные для использования в методе: Евклидово расстояние, квадрат Евклидова расстояния, расстояние Чебышева. Для построения минимального остовного дерева используется алгоритм Прима, т.к. он имеет наименьшую аналитическую сложность по сравнению с алгоритмами Крускала и Борувки [6].

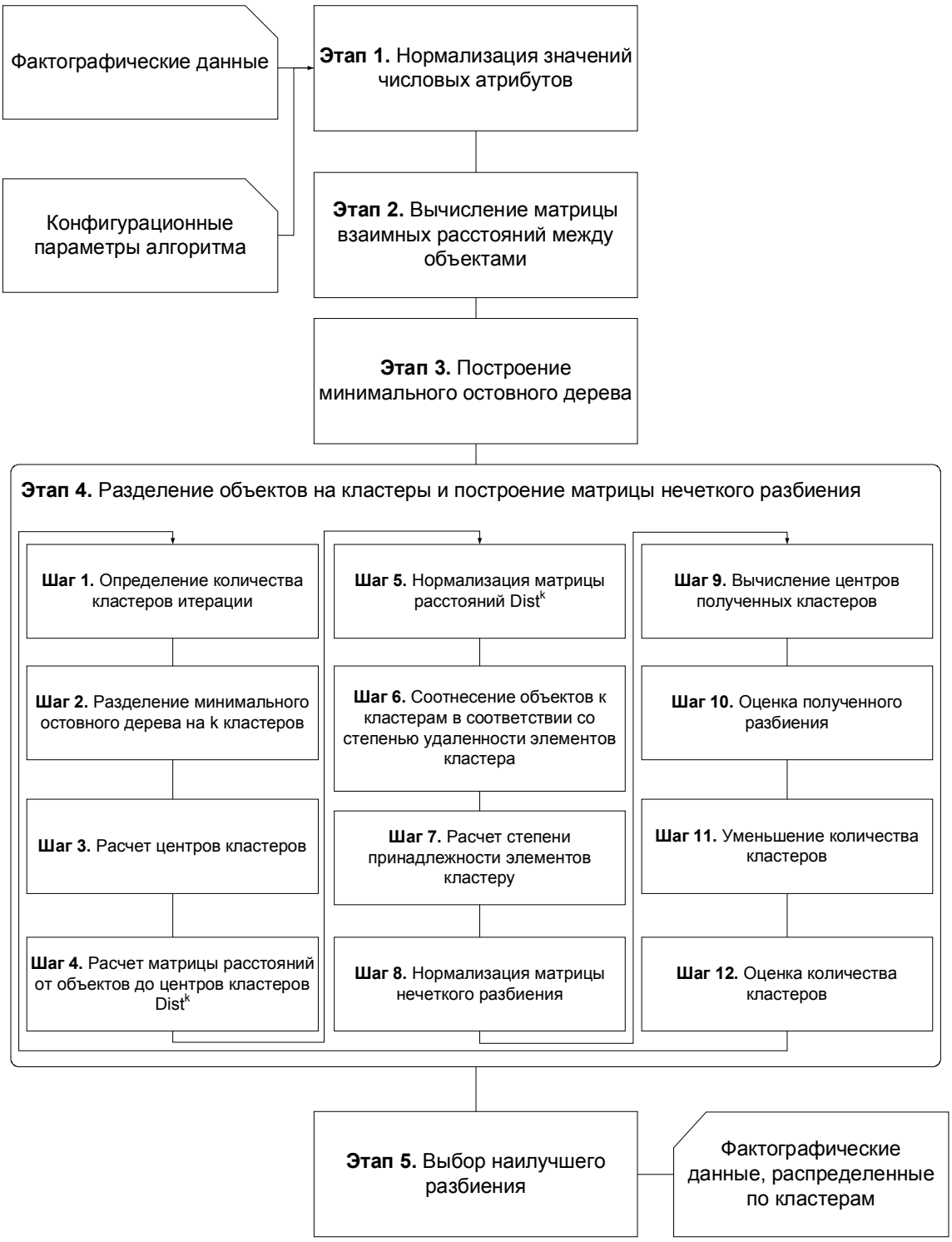


Рис. 1. Метод ADAKL

Оценка качества в методе ADAKL выполняется на основе локального критерия с использованием полученных центров кластеров:

$$O^k = \frac{\sum\limits_{i=\overline{1,k}} \dfrac{\left|V_i^{k'}\right| * \sum\limits_{j=1}^{m} \mu^{p}_{ij} * \left\| V_i^{k'} - u_j \right\|_{u_j \in V_i^{k'}}}{\underset{i \neq j}{Min}\left(\left\| V_i^{k'} - u_j \right\|_{u_j \in V_i^{k'}} \right) * Max\left(\left\| V_i^{k'} - u_j \right\|_{u_j \in V_i^{k'}} \right) * \sum\limits_{j=1}^{m} \left\| V_i^{k'} - u_j \right\|_{u_j \in V_i^{k'}} * k}}{m * k^2},$$

где k – количество кластеров;

m – количество объектов кластеризации;

$|V_i^{k'}|$ – количество элементов в кластере i;

$\mu_{ij}^{\ p}$ – степень принадлежности i-го объекта к j-му кластеру;

p – размазанность кластеров;

$\| V_i^{k'} - u_j \| = Metric(V_i^{k'}, u_j)$ – расстояние от центра кластера i до элемента u_j; $u_j \in V_i^{k'}$ – отражение условия о принадлежности элемента кластеру.

Данная оценка нацелена на выделение кластеров с наименьшими взаимными расстояниями и наибольшим количеством элементов в кластере по отношению к общему количеству кластеров, стремится к уменьшению количества итоговых кластеров и нацелена на минимизацию взаимных расстояний между полученным центром кластера и элементами с учетом степени принадлежности.

Предложенный метод обладает следующими достоинствами:

- двухэтапная кластеризация, которая позволяет выделить большее количество закономерностей;
- способен работать с лингвистическими атрибутами объектов, позволяя решить проблему использования экспертных оценок и текстовых атрибутов объектов;
- использует весовые коэффициенты для анализируемых атрибутов, позволяя не менять результирующий набор данных и работать со всем массивом, варьируя влиянием атрибута на результат анализа;
- использует степень удаленности объектов/элементов, позволяя соотносить объекты по кластерам при разделении на основе вычисленного расстояния;
- использует размазанность кластера, которая позволяет определять четкость получаемых границ кластеров;
- использует критерий оценки разбиения на кластеры, который учитывает требования и специфику предметной области.

Вместе с тем предложенный метод обладает квадратичной зависимостью аналитической сложности алгоритма от количества исходных

данных по объектам кластеризации, что существенно увеличивает временные затраты при регулярном появлении новых данных и повторной кластеризации.

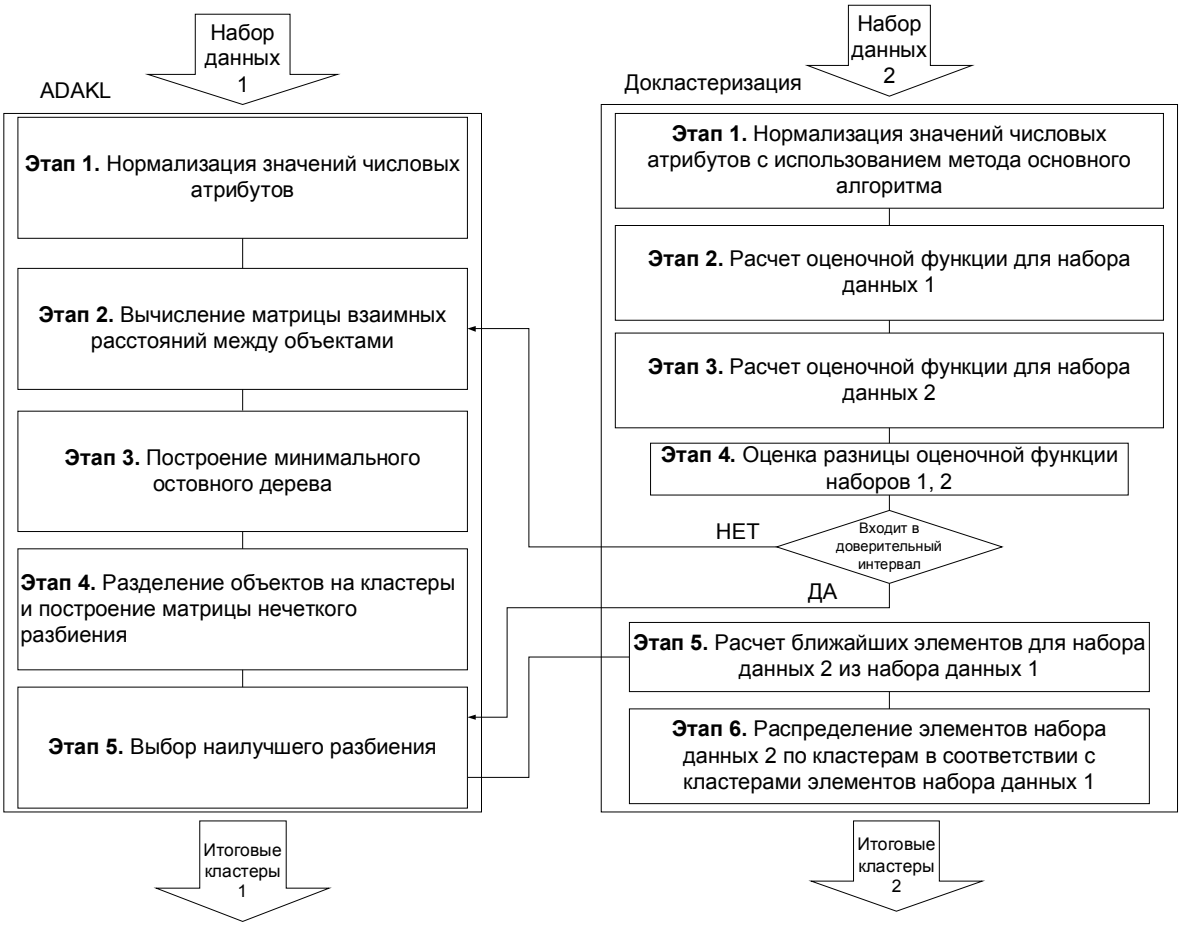


Рис. 2. Докластеризация дополнительного набора данных

Частично преодолеть этот недостаток можно за счет специальной процедуры докластеризации, которая определяет необходимость повторного запуска исследования полного массива данных и, в случае отсутствия признаков появления новых значимых групп объектов, осуществляет распределение новых (расширяющих) объектов по имеющимся кластерам. Для расширения исходных данных в процессе проведения анализа необходимо произвести дополнительное исследование добавляемых данных (рис. 2).

Необходимость в докластеризации подтверждается результатами эмпирических исследований, по результатам которых выявлено, что наиболее трудоемким этапом метода является построение минимального остовного дерева [5]. Выполнение дополнительного исследования при расширении исходных данных позволяет значительно сократить временные затраты по анализу данных за счет распределения расширяющих

объектов по имеющимся кластерам в случае подобности исходных данных в наборах 1, 2.

Для оценки работоспособности ADAKL в сравнении с другими алгоритмами были проведены три основных и одна дополнительная серии экспериментов:

- выделение секторов инвестирования на основе анализа показателей финансовых инструментов;
- выделение групп клиентов на основе статистических данных о деятельности клиентов за период;
- выявление категорий финансовых инструментов для оценки эффективности операций;
- выделение классов автомобилей на основе данных о максимальной скорости, цвете кузова, воздушном сопротивлении, массе.

Исследование производилось на трех методах (самоорганизующиеся карты Кохонена, алгоритм k-средних и разработанный метод ADAKL) с помощью аналитической платформы Deductor и разработанного программного решения, в котором реализован метод ADAKL. По результатам исследования составлена сводная таблица с усредненными оценками разбиений (табл. 1).

Таблица 1

Средневзвешенная оценка разбиений

Оценка / Метод	Средневзвешенная оценка разбиения	Средневзвешенная оценка разбиения с заданным количеством кластеров (без учета лингвистических атрибутов)	Средневзвешенная оценка разбиения с заданным количеством кластеров (с учетом лингвистических атрибутов)	Итоговая оценка
Карты Кохонена	0.7913	0.9150	0.9237	0.8767
k–средних	–	0.8232	–	0.8232
ADAKL	0.9762	0.9981	0.9990	0.9911

В соответствии с полученной итоговой оценкой наилучшее разбиение на исследованных массивах по сериям экспериментов получено с применением разработанного метода ADAKL. Проведенные эксперименты подтвердили, что использование интеграции методов кластеризации (многоэтапная кластеризация) улучшает качество выявления знаний в сравнении с одноэтапными методами, а также то, что превосходство разработанного метода достигается использованием математического аппарата нечеткой логики и внутренних словарей системы при определении информационных расстояний между объектами.

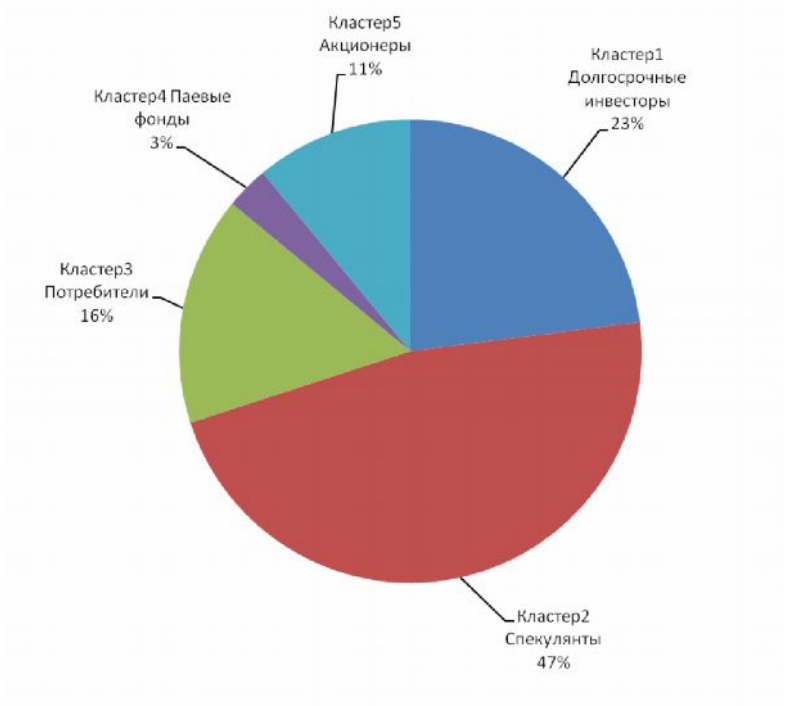


Рис. 3. Распределение клиентов по группам

На основе метода ADAKL были выделены группы клиентов и определение их доли от общего количества клиентов (рис. 3). Последующий анализ экономических показателей полученных групп объектов позволил дать названия кластерам, и разработать более целевую, направленную на конкретную клиентскую группу тарифную политику, а также предложить им более выгодные условия по совершаемым видам операций, увеличив количество этих операций и объем комиссионных сборов, что положительно повлияет на доходность данного направления деятельности кредитной организации. Дополнительная информацию о методике адаптивной кластеризации представлена в публикации [5] и на сайте научно-образовательного кластера CLAIM (http://philippovich.ru).

Литература

1. Прытин Д. Крупнейшие брокеры России // Источник: http://rating.rbc.ru.
2. Нейский И.М. Классификация и сравнение методов кластеризации // Интеллектуальные технологии и системы. Сборник учебно-методических работ и статей аспирантов и студентов. – М.: Изд-во ООО “Эликс +”, 2008. – Выпуск 8.
3. Speer N., Merz P., Spieth C., Zell A. Clustering Gene Expression Data with Memetic Algorithms based on Minimum Spanning Trees. // University of Tubingen, Center for Bioinformatics. Источник: fs.informatik.uni-tuebingen.de.

4. Штовба С. Д. Введение в теорию нечетких множеств и нечеткую логику // Источник: matlab.exponenta.ru.
5. Нейский, И.М., Филиппович, А.Ю. Методика адаптивной кластеризации фактографических данных на основе интеграции алгоритмов MST и Fuzzy C-means // Проблемы полиграфии и издательского дела. – М.: Изд-во МГУП, 2009. – №3.
6. Рыбаков Г. Построение минимального остовного дерева (алгоритмы Крускала, Прима, Борувки). – 2005. Источник: http://rain.ifmo.ru.

ПРИМЕНЕНИЕ АНАЛИТИЧЕСКОЙ ПЛАТФОРМЫ DEDUCTOR ДЛЯ АНАЛИЗА ПРЕЦЕДЕНТОВ ДИАГНОСТИКИ КРАНОВ МОСТОВОГО ТИПА

Климчук С.А., *аспирант Восточноукраинского национального университета им. В. Даля, г. Луганск, Украина*

Введение. Как показал опыт разработки системы поддержки принятия решений (СППР) технической диагностики, использование только двух источников знаний, а именно: экспертов и проблемно-ориентированных естественно-языковых текстов, часто приводит к неполноте извлекаемых знаний.

В условиях приобретения знаний неполнота связана, в основном, с тем, что эксперт не знает (не отметил, либо забыл отметить) какой-либо факт, необходимый для решения задачи. В этом случае возможны следующие альтернативы преодоления неполноты: либо проведение нескольких сеансов приобретения знаний с одним и тем же экспертом и сравнение полученных результатов, либо привлечение нескольких экспертов и корреляция их мнений, а также использование технологии извлечения знаний из баз данных – Data Mining.

Целью данной работы является применение технологии Data Mining в рамках аналитической платформы (АП) Deductor для разработки СППР при решении одной из востребованных неформализованных задач – задачи технической диагностики кранов мостового типа.

Описание возможностей АП Deductor. Аналитическая платформа Deductor (разработчик – компания BaseGroup Labs) – программа, реализующая функции импорта, обработки, визуализации и экспорта данных. Составной компонент платформы – приложение *Deductor Studio* может

функционировать и без хранилища данных, получая информацию из любых других источников, но наиболее оптимальным является их совместное использование. В *Deductor Studio* включен полный набор механизмов, позволяющий получить информацию из произвольного источника данных, провести весь цикл обработки (очистку, трансформацию данных, построение моделей), отобразить полученные результаты наиболее удобным образом (OLAP, диаграммы, деревья…) и экспортировать результаты на сторону.

Описание метода решения задачи. Сети, называемые картами Кохонена, – это одна из разновидностей нейронных сетей, которые используют неконтролируемое обучение. При таком обучении обучающее множество состоит лишь из значений входных переменных, в процессе обучения нет сравнивания выходов нейронов с эталонными значениями. Можно сказать, что такая сеть учится понимать структуру данных.

Наиболее распространенное применение сетей Кохонена – решение задачи классификации без учителя, т.е. кластеризации. При такой постановке задачи нам дан набор объектов, каждому из которых сопоставлена строка таблицы (вектор значений признаков). Требуется разбить исходное множество на классы, т.е. для каждого объекта найти класс, к которому он принадлежит. В результате получения новой информации о классах возможна коррекция существующих правил классификации объектов.

Постановка задачи. Пусть имеется база данных прецедентов диагностики, содержащая сведения об обследованиях мостового крана за весь срок его эксплуатации. Необходимо выявить влияние таких факторов, как остаточный прогиб в вертикальной плоскости, искривление балок и ферм в плане, наличие металла с ударной вязкостью меньше 2 кгс·м/см^2 на изменение коррозии.

Таким образом, входными параметрами будут:

- остаточный прогиб;
- искривление балок и ферм;
- наличие металла с ударной вязкостью.

Выходной параметр – изменение коррозии в%:

- свыше 20% – не допускается;
- от 10 до 20% – допускается, но с перерасчетом грузоподъемности;
- менее 10% – допускается.

Решение задачи средствами АП Deductor. Необходимо провести кластеризацию параметров, т.е. выделить однородные группы на основе показателей из базы данных.

Сначала импортируем данные из файла в *Deductor Studio*. Затем запускаем мастер обработки и выбираем из списка метод обработки *Карта Кохонена*. Далее следует настроить назначения столбцов, т.е. для каждого столбца выбрать одно из назначений: входное, выходное, не используется и информационное. Укажем столбцам *Остат. прогиб*, *Искривление*, *Ударн. вязкость* назначение *Входной*. *Выходной* назначаем для столбца *Коррозия* (рис. 1).

Следующий шаг предлагает разбить исходное множество на обучающее и тестовое. По умолчанию предлагаются следующие пропорции: обучающее – 95%, тестовое – 5%.

На следующем шаге предлагается настроить параметры карты: количество ячеек по оси *X* (по умолчанию – 16) и по оси *Y* (по умолчанию – 12), а также их форму (шестиугольную или прямоугольную).

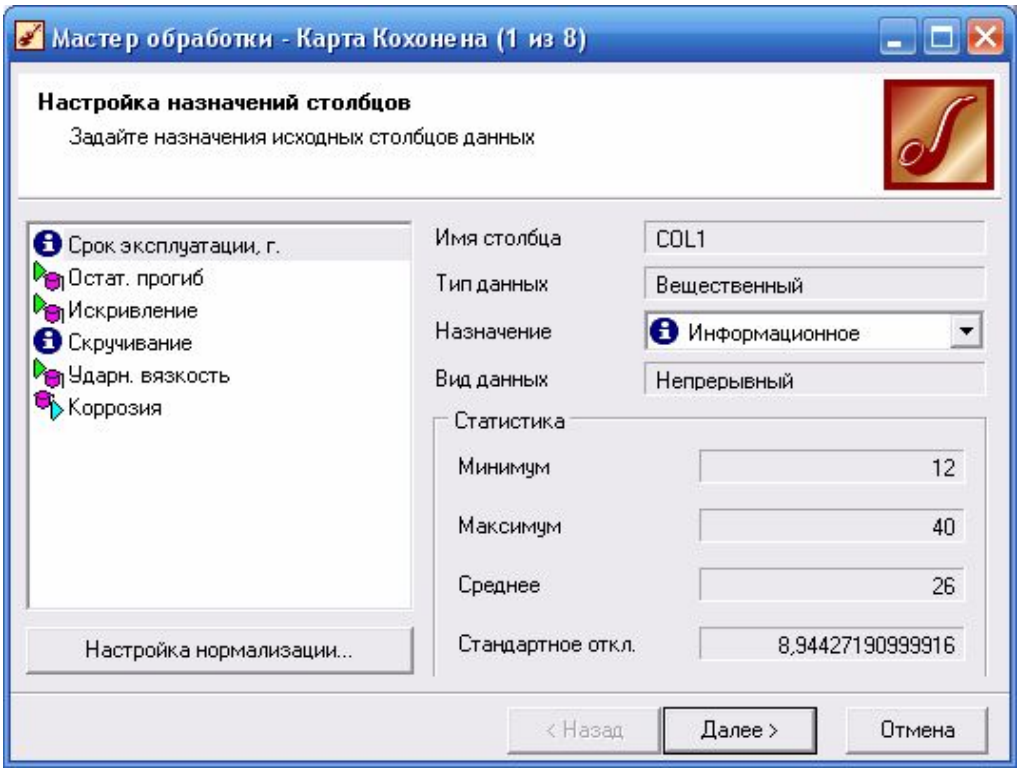


Рис. 1. Настройка назначений столбцов

На шаге «Настройка параметров остановки обучения» устанавливаем параметры остановки обучения и устанавливаем эпоху (по умолчанию – 500), по достижению которой обучение будет прекращено.

На следующем шаге, представленном на рис. 2, настраиваются другие параметры обучения: способ начальной инициализации, тип функции соседства. Поскольку нам неизвестно количество кластеров, выберем автоматическое определение их количества.

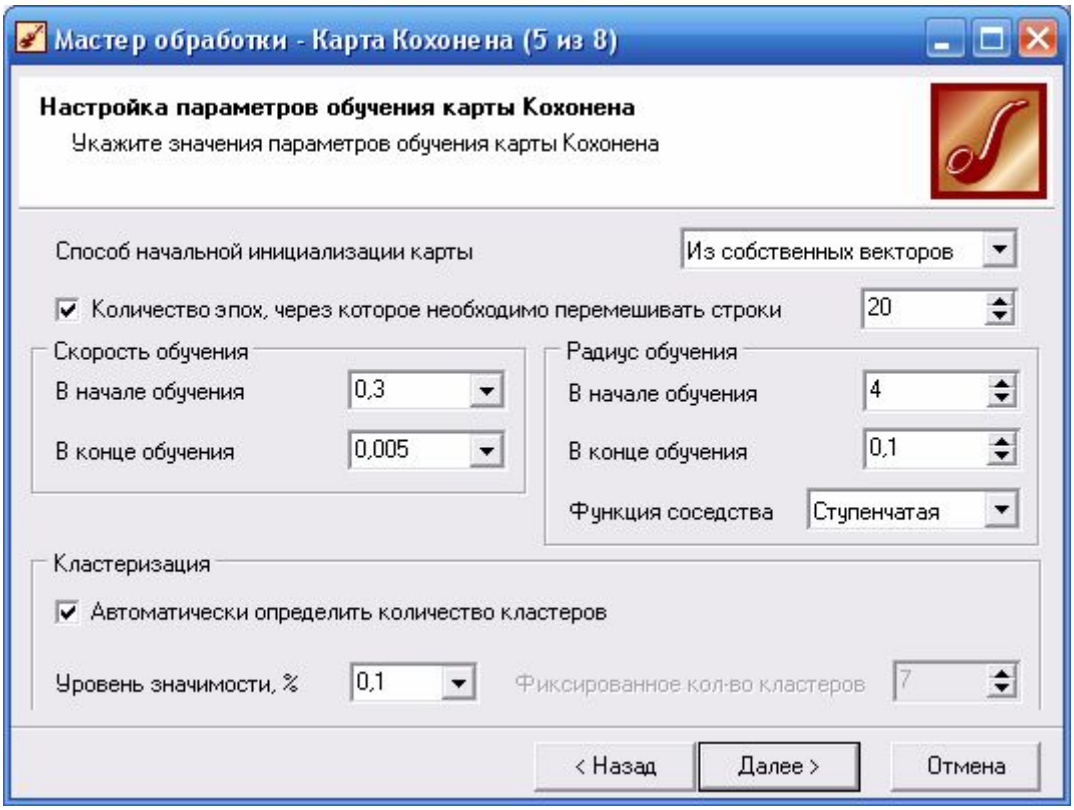


Рис. 2. Настройка параметров обучения

Далее запускаем процесс обучения сети – необходимо нажать на кнопку *Пуск* и дождаться окончания процесса обучения. Во время обучения можем наблюдать изменение количества распознанных примеров и текущие значения ошибок.

По окончании обучения в списке способов отображения данных выберем «Карта Кохонена», «Профили кластеров» и визуализатор «Что-если». На последнем шаге настраиваем отображения карты Кохонена.

На рис. 3 приведена иллюстрация карт входов и выходов, последняя – это карта кластеров. Здесь мы видим несколько карт входов и сформированные кластеры, каждый из которых выделен отдельным цветом.

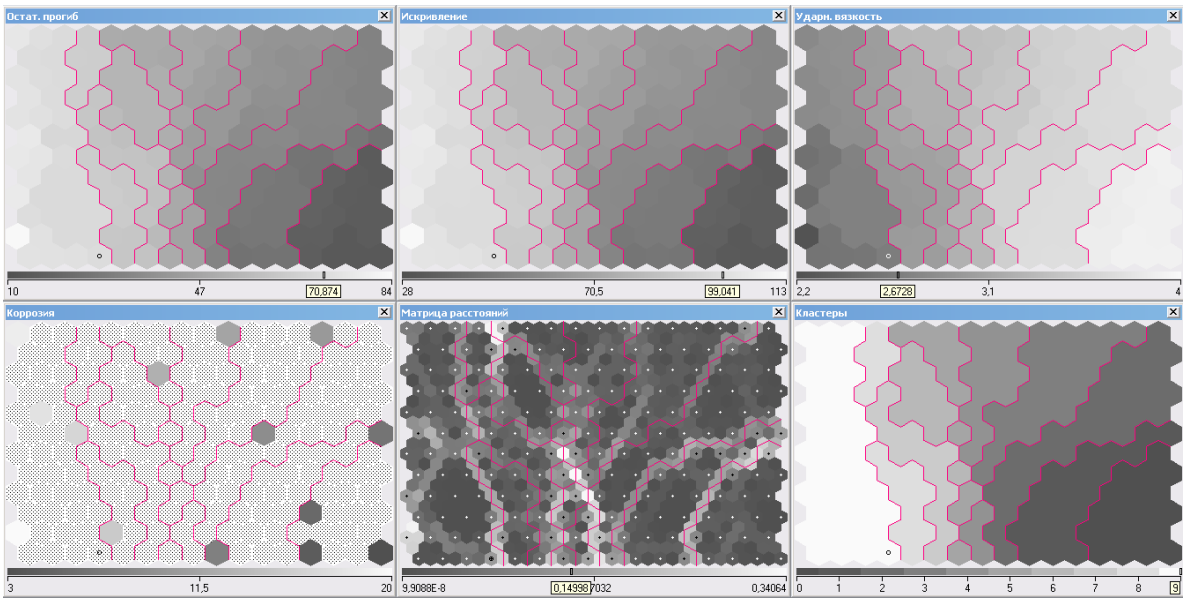


Рис. 3. Карты входов и выходов

Анализ полученных результатов. В результате решения задачи получено 9 кластеров.

При анализе карт входов рекомендуют использовать сразу несколько карт. Исследуем фрагмент карты, состоящий из карт трех входов, который приведен на рис. 3.

На карте *Остаточный прогиб* выделяем область с наибольшим значением показателя. Далее имеет смысл изучить эти же ячейки на других картах. Они расположены также в правом нижнем углу и относятся к 9 кластеру, характеризующемуся самыми высокими показателями коррозии.

Также мы можем определить, например, такую характеристику: кластер, расположенный в правом верхнем углу, характеризуется высокими значениями показателя искривления.

Для нахождения конкретного объекта на карте необходимо нажать правой кнопкой мыши на исследуемом объекте и выбрать пункт «Найти ячейку на карте». В результате мы можем видеть как сам объект, так и значение того измерения, которое мы просматриваем. Таким образом, мы можем оценить положение анализируемого объекта, а также сравнить его с другими объектами.

Данные по обследованию крана были классифицированы на 9 групп, для каждой из которых возможно определение конкретных характеристик, исходя из раскраски соответствующих показателей.

Таким образом, карты Кохонена позволяют упростить многомерную структуру, их можно считать одним из методов проецирования многомерного пространства в пространство с более низкой размерностью. Интенсивность цвета в определенной точке карты определяется данны-

ми, которые туда попали: ячейки с минимальными значениями изображаются черным цветом (в цветной палитре – темно-синим), ячейки с максимальными значениями – белым (в цветной палитре – красным).

Другое принципиальное отличие карт Кохонена от других моделей нейронных сетей – иной подход к обучению, а именно – неуправляемое или неконтролируемое обучение. Этот тип обучения позволяет данным обучающей выборки содержать значения только входных переменных. Сеть Кохонена учится понимать саму структуру данных и решает задачи кластеризации.

Заключение. Важным этапом в процессе Data Mining является предварительная подготовка данных, в том числе их очистка. От качества подготовленных данных будут зависеть результаты всего процесса.

В процессе построения и выбора модели Data Mining следует пробовать использовать различные методы и алгоритмы, а также их сочетания. При отсутствии опыта использования методов Data Mining лучше начинать с более простых, поддающихся интерпретации моделей. Далее можно постепенно усложнять модели, т.е. использовать более сложные методы. Не следует требовать от модели абсолютной точности, модель можно начинать использовать при получении первых приемлемых результатов.

Следует помнить, что процесс Data Mining является итеративным. При невозможности получения результатов, которые эксперт предметной области считает приемлемыми, необходимо вернуться на один из предыдущих этапов процесса.

Рассмотренный пример демонстрирует большие возможности аналитической платформы Deductor при решении широкого спектра задач, которые связаны с обработкой структурированных (табличных) данных. Он предоставляет аналитикам инструментальные средства, необходимые для решения самых разнообразных аналитических задач, начиная от всевозможной аналитической отчётности и заканчивая созданием на его базе СППР по технической диагностике кранов мостового типа.

Литература

1. Deductor [Электронный ресурс]. – Электрон. дан. – 20.08.2010. – Режим доступа: http://basegroup.ru/deductor. – Загл. с экрана.
2. Барсегян А.А. Технологии анализа данных: Data Mining, Visual Mining, Text Mining, OLAP / А.А. Барсегян, М.С. Куприянов, В.В. Степаненко, И.И. Холод. – 2-е изд., перераб. и доп. – СПб.: БХВ – Петербург, 2007. – 384 с.
3. Хайкин С. Нейронные сети: полный курс / С. Хайкин. – 2-е изд.: Пер. с англ. – М.: Издательский дом «Вильямс», 2006. – 1104 с.

4. Климчук С.А. Система поддержки принятия решений при диагностировании кранов мостового типа на основе прецедентов / С.А. Климчук / Системный анализ и информационные технологии: материалы 12-й Международной научно-технической конференции SAIT 2010, Киев, 25-29 мая 2010 г. / УНК «ИПСА» НТУУ «КПИ». – К.: УНК «ИПСА» НТУУ «КПИ», 2010. – С. 260.

МЕТОД КЛАСТЕРИЗАЦИИ ТЕКСТОВ HTECL

Стулов В.В., *магистр техники и технологии,* ***Филиппович А.Ю.,*** *доцент, Московский государственный технический университет им. Н.Э. Баумана*

Введение. Кластеризация корпуса текстов есть разделение его на группы (кластеры) семантически подобных между собой текстов. Данная процедура может быть использована для анализа и поиска в больших текстовых коллекциях, таких как web-страницы. Кластеризация результатов поиска позволяет как выявлять среди релевантных документов группы наиболее релевантных, так и семантически группировать релевантные группы в тематические кластеры, а также образовывать иерархию кластеров (таксономию). Кластеризация результатов поиска применяется в Интернет-поисковиках *Nigma* и *Clusty*, однако выделяемые поисковыми машинами кластеры не являются тематическими.

В статье представлен метод дивизимной иерархической кластеризации HTECL. Ключевыми особенностями метода являются: использование векторной модели текста, ориентированной на семантику, отсутствие предопределенной формы получаемой таксономии, применение нечеткой сети Кохонена для кластеризации на каждом уровне иерархии, отсутствие необходимости задания числа кластеров на каждом уровне иерархии, кластеризация ограниченного набора элементов, устранение ошибок классификации за счет введения процедуры слияния кластеров.

Модели представления текста. Существует множество подходов к формализации корпуса текстов и соответствующих им алгоритмов кластеризации. Алгоритм представленный в [2] использует множество частотных слов всего корпуса текстов для дальнейшего распределения отдельных текстов по кластерам, руководствуясь принципом минимума пересечения кластеров. Основным преимуществом используемого под-

хода является быстрота кластеризации, однако точность кластеризации является относительно малой. Алгоритм [3] использует в своей работе суффиксное дерево корпуса текстов.

Основным преимуществом алгоритмов, использующих модель суффиксного дерева, является быстрота кластеризации. Форма суффиксного дерева зависит от порядка слов в предложении, что отрицательно сказывается на точности кластеризации. Кроме того, данная структура не ориентирована на выявление и использование семантики текстов. Согласно источникам [2], [5] получающиеся в результате работы алгоритмов кластеры определяются некачественно.

Решением проблемы точности кластеризации является использование структуры суффиксного дерева синтаксем в качестве модели корпуса текстов. В данной модели метками дуг дерева являются не отдельные слова, а синтаксемы предложений либо их части. Однако данных подход имеет недостаток связанный с трудоемкостью определения функции слова в предложении. Эта процедура требует построения специальных электронных словарей.

Большинство существующих алгоритмов кластеризации используют в своей работе формальные модели отдельных текстов, а не целого корпуса. В качестве основной модели текста выступает векторная модель. В этой модели текст представляет собой точку в многомерном пространстве признаков. Под признаком понимается любое слово, присутствующее в конкретном тексте, либо его основная форма (лемма). Самой простой моделью является бинарный вектор:

$$d = (t_1,...,t_m),\ \forall i \equiv \overline{1,m}, t_i \in \{0,1\}, \qquad (1)$$

указывает, содержит ли текст слово, ассоциированное с признаком *i*. В данном виде векторная модель практически никогда не используется, так как в ней не учитывается относительный семантический вес признаков. В большинстве случаев вес признака учитывается посредством вычисления относительной частоты данного признака либо внутри документа, либо по всему кластеризуемому корпусу. В первом случае значение $t_i = \frac{Nt_i}{N}$, где Nt_i – число раз, которое признак *i* встречается в тексте, N – общее число слов текста. Во втором случае (так называемая мера TF-IDF) $t_i = \frac{Nt_i}{N} \cdot \log \frac{D}{D_i}$, где D – общее число документов корпуса, D_i – число документов, в которых встречается признак *i*. Большой вес в TF-IDF получают слова (леммы) с высокой частотой в пределах конкретного документа и с низкой частотой употреблений в других документах. Большую частоту как в модели частотного словника так и, зачастую, в модели

TF-IDF получают семантически не значимые, однако высоко частотные предлоги и союзы. Поэтому алгоритмы, использующие данные модели исключают из пространства признаков данные слова.

Основной проблемой подходов формирования вектора текста на основе частот признаков заключается в том, что эти модели не учитывают взаимные связи между словами в предложениях, которые отражают его семантику. Для примера рассмотрим два предложения: *1. На краю дороги стоял дуб. 2. Это был огромный, в два обхвата дуб, с обломанными давно, видно, суками и с обломанной корой, заросшей старыми болячками.* Видим, что семантический вес слова "дуб" выше во втором предложении, чем в первом, за счет его богатого описания.

Для учета семантического веса слов в предложении разработанный метод использует модель векторного представления текста, предложенную в [4]. Данная модель ранжирования весов признаков использует предположение от том, *что чем больше связей у слова в дереве синтаксического разбора текста, тем больше его вес в данном предложении*.

Приведем алгоритм формирования векторов текстов предлагаемого метода кластеризации.

1. Морфологический и синтаксический анализ текста. Выполняется с помощью программы *Cognitive Dwarf* [8]. На данном этапе лемматизируются слова, содержащиеся в тексте и строится деревья синтаксического разбора предложений.
2. Строится матрица M размера n x n, где n – число слов в текста за исключением союзов и предлогов. Каждый элемент матрицы m_{ij} равен числу связей вида *термин*$_i$ → *термин*$_j$ и *термин*$_j$ → *термин*$_i$ в дереве синтаксического разбора текста.
3. Вектор модели $V = \{v_i\}$, где $v_i = \sum_j m_{ij}$.

В качестве примера рассмотрим модель текста, состоящего из следующего предложения: *Применение данной модели позволило достичь высокого качества кластеризации*.

На рис. 1 изображено дерево синтаксического разбора текста, а также указаны получаемые веса терминов.

Применяемый подход имеет недостаток. Он заключается в том, что все связи графа считались одинаково значимыми и их веса были заданы значением 1. В [4] для задания силы связи между словами использовалось машинное обучение. Это позволило повысить точность кластеризации на ~8% (точность оценивалась как отношение числа документов, классифицированных верно к общему числу документов). Однако правомочность задания силы связи между словами вызывает сомнения. Ведь связи слов в предложении отличаются между собой своей семантикой и

связи двух одинаковых слов в различных предложениях могут быть семантически разными. Для строгой разметки графа синтаксического разбора текста целесообразнее использовать модель связей, приведенную в [7]. В этой модели каждое слово в предложении играет свою функцию.

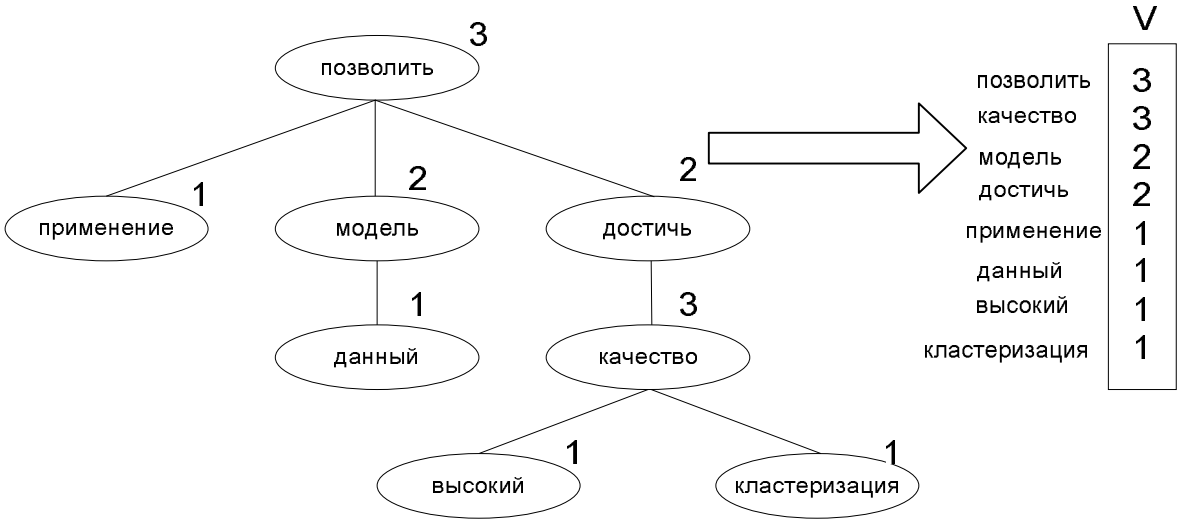


Рис. 1. Формирование векторного представления текста

Другим путем улучшения качества модели текста является подключение на этапе препроцессинга текста специального модуля с функцией тезауруса, который устраняет негативные последствия синонимии и омонимии, выполняет замену местоимений. Однако разработка такого модуля трудоемка, его внедрение уменьшает скорость всей процедуры кластеризации.

Мера сходства текстов. Разработанный метод кластеризации HTECL использует формулу (2) в качестве меры расстояния между векторами текстов:

$$D(\mathbf{x}_i, \mathbf{x}_j)\sqrt{\sum_k x_{ik}^2}\cdot\sqrt{\sum_k x_{jk}^2}-\sum_k (x_{ik}^{'}\cdot x_{jk}^{'}), \qquad (2)$$

где $x_{ik}^{'}$ и $x_{jk}^{'}$ – координаты нормализованных векторов $\mathbf{x}_i$ и $\mathbf{x}_j$.

Сокращение размерности пространства признаков. Существует два основных способа сокращения размерности - локальный метод и метод латентно-семантического анализа. HTECL в своей работе использует локальный метод – отдельные компоненты вектора документа (сумма v_i соответствующего терма по всем документам корпуса ниже определенного порога), исключаются из рассмотрения.

Определение числа кластеров. Определение числа кластеров является актуальной проблемой многих алгоритмов. Платформа *Deductor* использует алгоритм g-means, требующий задания пользователем «уровня значимости» для выявления количества кластеров.

Приведем теперь особенности разработанного метода нейросетевой кластеризации HTECL. Согласно алгоритму обучения нечеткой сети Кохонена, векторы синапсов, от итерации к итерации стремятся к центрам плотных скоплений точек многомерного пространства, которые находятся к ним ближе всего согласно выбранной меры. Рассмотрим случай постепенного увеличения количества нейронов. Пусть изначально число нейронов меньше чем число кластеров. В этом случае в результате проведения кластеризации центр кластера будет находиться между плотными скоплениями точек со смещением к скоплению большей мощности (рис 2).

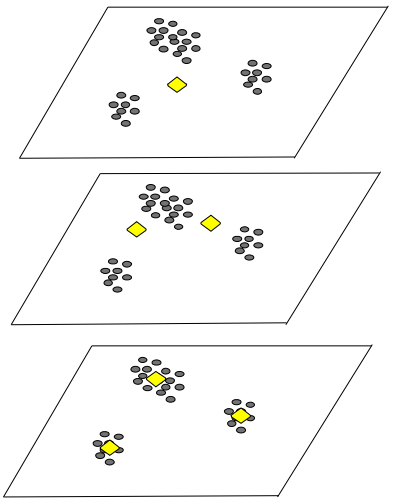


Рис. 2. Постепенное увеличение числа кластеров

Таким образом, значение функции $f(k) = \frac{\sum_{p=1}^{k} avg_i \mu_{ij}}{k}$, где $avg_i \mu_{ij}$ – средняя по кластеру принадлежность входящего в него вектора, будет монотонно возрастать при стремлении $k \to k'$, где k' – общее число плотных скоплений векторов. Вектор j считается принадлежащим кластеру i, если $\mu_{ij} > \frac{100\%}{k}$. При $k > k'$ значение функции $f(k)$ уменьшится, так как k возрастет, а средняя принадлежность уменьшится за счет того, что одно из скоплений будет поделено на два кластера (большая часть точек будет равноудалена от двух нейронов), рис 3.

Благодаря этому свойству алгоритма обучения сети Кохонена можно определить число кластеров (плотных скоплений) в обучающей выборке.

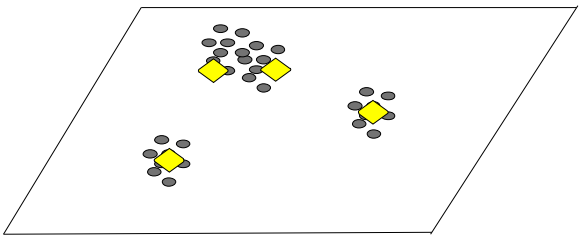


Рис. 3. $k > k'$

Таким образом, на каждом этапе иерархической кластеризации число кластеров первого уровня будем определять итеративно по точке экстремума функции $f(k)$. При этом $k = \overline{2, N/2}$ – в каждом кластере должно быть не менее двух текстов.

Сокращение времени кластеризации. Вычислительная сложность алгоритма обучения нечеткой сети Кохонена линейно зависит от числа векторов обучающей выборки. Для сокращения времени кластеризации на каждом уровне иерархии будем кластеризовать не всю выборку, а некоторую ее долю, мощность которой задается пользователем. Оставшиеся векторы подвергаются классификации. Затем кластеризация продолжается для каждого из определенных укрупненных кластеров, таким образом мы получаем вложенную иерархию кластеров.

Рассмотрим случаи, которые могут возникать в случае кластеризации неполной выборки.

1 случай–- кластеры определены верно.

$\forall x \in X, d(x, C) \le R$, где x –объекты кластера X, C – точка-центроид, R – радиус первичного кластера.

Данное событие можно отследить, если каждый из классифицированных векторов удален от центра кластера на расстояние, меньшее, чем радиус кластера. Под радиусом кластера понимается расстояние от центра кластера до наиболее удаленного от него вектора, принадлежащего этому кластеру.

2 случай – укрупненный кластер не является плотным скоплением векторов.

$\exists X_i, X_{i+1}, ..., X_j \in X$, где X – кластер высокого уровня, $X_i,\ldots,X_j$ – подкластера.

В данном случае в укрупненный кластер попадают несколько близких друг другу более мелких кластера, рис 4.

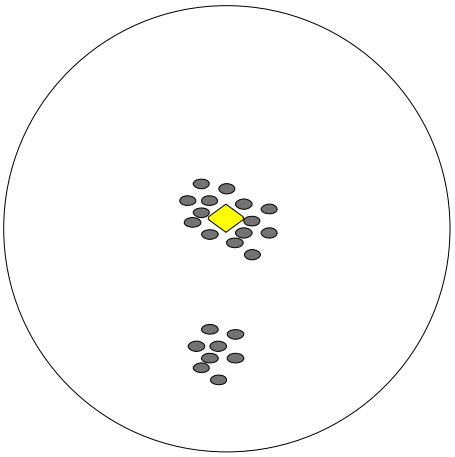


Рис. 4. Укрупнение кластеров

Данный случай является идеальным с точки зрения иерархической дивизимной кластеризации. На более высоком уровне мы определили крупный кластер, который распадется на два или более в результате дальнейшей кластеризации. Начальные положения центров кластеров при последующем запуске процедуры также будут выбираться случайным образом. Таким образом мы строим иерархию групп текстов.

3 случай - укрупненные кластеры разделяют между собой один или несколько кластеров.

$\exists x \in X, \exists y \in Y : x_i \cup y_i = Z$, где X, Y, Z – кластеры высокого уровня, x, y – объекты кластеров X и Y.

Данный случай приведен на рис. 5. При переходе на нижний уровень иерархии образуется множество новых кластеров. Для обнаружения кластера, более высокого уровня, необнаруженного ранее следует проверить все эти кластера на возможность слияния.

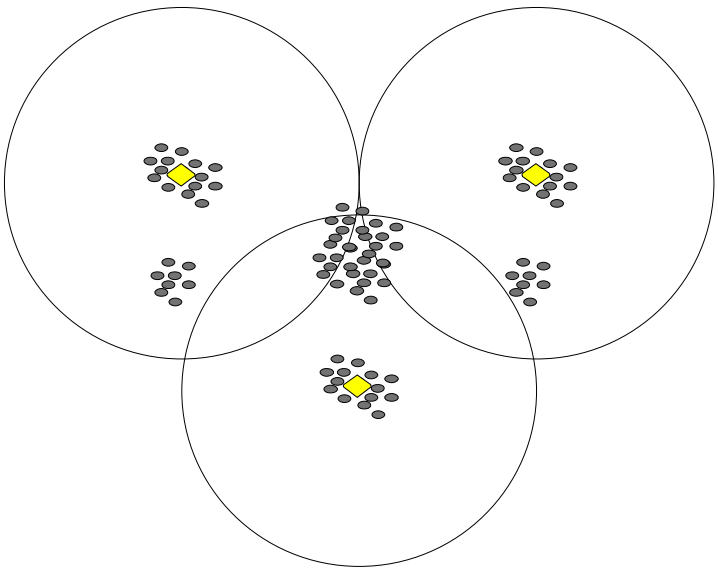


Рис. 5. Необнаруженный кластер

4 случай – укрупненные кластеры являются одним кластером.

$\forall x \in X : x \in Z \wedge \forall y \in Y : y \in Z$, где X, Y, Z – кластеры, x, y – объекты кластеров X и Y.

Этот случай иллюстрируется на рис. 6.

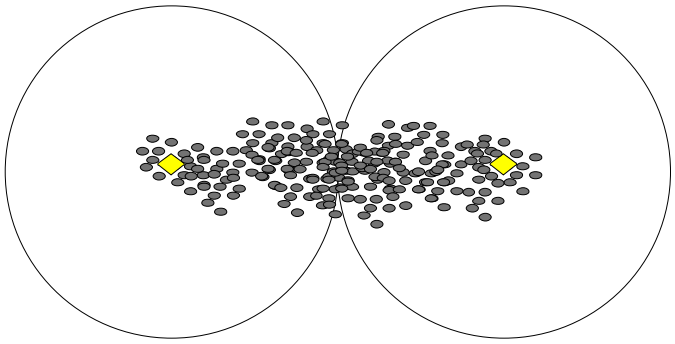


Рис. 6. Неверное разделение кластера

В данном случае каждый из укрупненных кластеров при последующей кластеризации не разделиться на подкластеры. Поэтому на каждом этапе кластеризации необходимо проверять возможность слияния полученных кластеров.

5 случай - укрупненный кластер содержит часть другого кластера.

$\exists x \in X : x \in Y$, где X, Y, Z – кластеры, x, y – объекты кластеров X и Y.

Этот случай иллюстрируется на рис. 7.

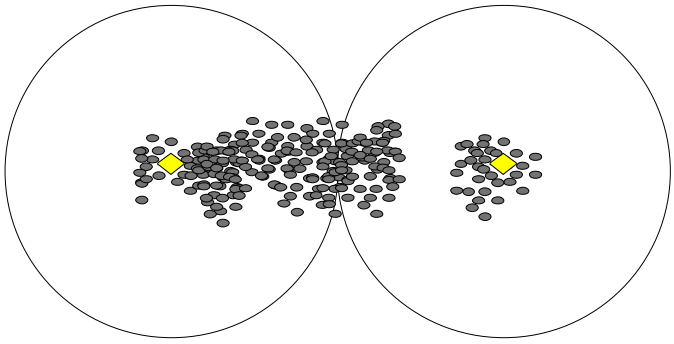


Рис. 7. Укрупненный кластер содержит часть другого кластера

Итого на каждом этапе кластеризации необходимо проверять полученные кластеры на необходимость их слияния, слияния подкластеров с подкластерами другого кластера, слияния подкластеров с другими кластерами высокого уровня.

Алгоритм. Приведем теперь формальное описание разработанного метода кластеризации текстов HTECL.

Этап 1. Препроцессинг текстовых файлов

Шаг 1. Морфологический анализ текстовых файлов.

Производится для удаления из рассмотрения семантически незначимых слов (предлоги и союзы) и лемматизации значимых.

Шаг 2. Синтаксический анализ текстовых файлов.

Строятся деревья синтаксического разбора предложений.

Шаг 3. Формирование векторов текстов.

Этап 2. Сокращение размерности пространства признаков

Этап 3. Кластеризация части векторов с помощью нечеткой сети Кохонена

Для определения количества кластеров используется алгоритм, приведенный в п.2.2.

Этап 4. Классификация оставшихся векторов

Этап 5.Проверка списка подлежащих докластеризации кластеров

Если список пуст, то переход к этапу 6. Иначе, переход к этапу 8.

Этап 6. Процедура слияния кластеров

Процедура описана в [10].

Этап 7. Определение списка кластеров, подлежащих докластеризации

В случае если количество уровней иерархии не ограничено пользователем, в список заносятся все листьевые кластера с количеством элементов большим двух.

Этап 8. Извлечение кластера на докластеризацию

Из списка кластеров, подлежащих докластеризации, выбирается очередной кластер. Переход на этап 2. Если список пуст, переход на этап 9.

Этап 9. Конец

Для оценки метода была осуществлена серия испытаний. В качестве корпуса текстов был выбран Reuters-21578. Точность кластеризации составила ~74%, что на ~10% выше точности алгоритма k-means на том же наборе исходных данных.

Литература

1. Feldman R., Sanger J. The Text Mining Handbook. – Cambridge University Press, 2006.
2. Beil F., Ester M., Xu X. Frequent Term-Based Text Clustering // Proc. 8th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD ‘2002), Edmonton, Alberta, Canada, 2002.
3. Zamir O., Etzioni O. Web Document Clustering: A Feasability Demonstration // Proc. ACM SIGIR 98, 1998, P. 46-54.
4. Choudhary B., Bhattacharyya P. Text Clustering using Semantics // The Eleventh International World Wide Web Conference, 2002.

5. А.М. Андреев, Д.В. Березкин, В.В. Морозов, К.В. Симаков: Метод кластеризации документов текстовых коллекций и синтеза аннотаций кластеров // Электронные библиотеки: перспективные методы и технологии, электронные коллекции: Десятая Всерос. научн. конф., 2008.
6. Magnus Rosell: Introduction to Information Retrieval and Text Clustering, KTH CSC, 2006.
7. Г.С. Осипов, И.В. Смирнов, И.А. Тихомиров: "Реляционно-ситуационный метод поиска и анализа текстов и его приложения", Искусственный интеллект и принятие решений 2008 / 02
8. Описание программного пакета синтаксического разбора и машинного перевода, Cognitive Technologies, Ltd., 2006.
9. Нейский И.М.: Классификация и сравнение методов кластеризации. // Интеллектуальные технологии и системы. Сборник учебно-методических работ и статей аспирантов и студентов. Выпуск 8. – М.: Изд-во ООО “Эликс +”, 2008. – С. 111-122.
10. Стулов В.В. Методика нейросетевой кластеризации корпуса текстов. IV всероссийская студенческая научно-техническая конференция “Прикладная информатика и математическое моделирование”, 19-20 мая 2010 года, МГУП [в печати].

ПОПУЛЯЦИОННО-ГЕНЕТИЧЕСКИЕ МЕТОДЫ РЕШЕНИЯ ЗАДАЧ ДИСКРЕТНОЙ ОПТИМИЗАЦИИ ПОВЫШЕННОЙ РАЗМЕРНОСТИ[1]

***Паклин Н.Б.**, доцент Рязанского государственного университета им. С.А.Есенина, г. Рязань, **Крепышев Д.А.**, старший преподаватель Кубанского государственного аграрного университета, г. Краснодар*

Введение. В практике экономико-математического моделирования часто встречаются задачи дискретного программирования повышенной размерности (тысячи переменных). Задачи подобного рода встречаются в сельском хозяйстве, промышленности, управлении, однако, решаются не удовлетворительно. Программные комплексы, базирующиеся в основном

[1] Работа поддержана грантом РФФИ №10-01-00070-а

на модифицированных симплекс-методах, плохо приспособлены для решения такого рода задач, хотя их разработчики добились определенных успехов [1].

Поэтому актуальным представляется разработка и исследование методов, направленных на устранение недостатков модифицированных симплекс-методов. Большое поле для этой деятельности в последние годы предоставили *эволюционные* алгоритмы, в особенности *популяционные* и *ройные*. Из первых следует выделить *генетические алгоритмы* (ГА), обсуждению которых применительно к задачам дискретной оптимизации повышенной размерности посвящена эта статья.

Задачи повышенной размерности. В работах исследователей можно встретить следующее условное деление задач по их размерности [5, 6]: малоразмерные задачи (с матрицей от 1000x1000 до 4000x4000), задачи повышенной размерности (от 4000x4000 до 6000x6000) и задачи большой размерности (от 6000x6000 до 10000x10000). Конечно, такое деление меняется с ростом вычислительных возможностей компьютерной техники. Однако можно сформулировать общее правило отнесения задачи к классу повышенной размерности: когда стандартные алгоритмы и программные средства не позволяют получить за ограниченное время приемлемое решение. Ограничение по времени может быть продиктовано максимально допустимым временем отклика в системе поддержки принятия решений, либо разумными пределами ожидания (несколько минут, часов). В этом случае наблюдается противоречие между экспоненциальным ростом вычислительной сложности, с одной стороны, и постоянно растущими ограничениями по времени решения, с другой стороны. Сегодня очевидно, что без эвристических, гибридных методов данное противоречие не разрешить.

Перечислим наиболее распространенные прикладные задачи дискретной оптимизации: задача об укладке рюкзака, задача коммивояжера, одномерный раскрой листовых материалов разных размеров, задача о покрытии множества системой его подмножеств, оптимизация структуры атомного кластера, транспортные задачи, составление планов и расписаний. Их математические постановки хорошо описаны в соответствующей литературе [1, 3, 4, 7]. К традиционным методам решения этих задач относят экспоненциальные алгоритмы, которые строятся на основе свойств целевой функции и ограничений. Наиболее известны [8]: симплекс-метод для решения задач целочисленной оптимизации с линейными ограничениями; группа методов последовательного анализа и отсеивания вариантов, который является развитием метода «ветвей и границ» для задачи дискретной оптимизации с неубывающими целевыми функциями, и позволяет по анализу некоторого числа вариантов отсеивать

большее число, последовательно уменьшая множество вариантов до размеров, удовлетворительных для использования прямого перебора. Это так называемые *точные* методы. Параллельно развивались *приближенные* методы: локальная оптимизация, эвристические процедуры, максимально учитывающие специфику решаемых задач, метод вектора спада, метод направляющих окрестностей, методы случайного поиска и другие [8].

Популяционно-генетические методы. В последние годы накоплены десятки тысяч исследовательских работ посвященные эволюционным вычислениям. Их разновидность, генетический алгоритм (ГА), относится к классу методов случайного направленного поиска и, могут быть отнесены к классу приближенных методов дискретной оптимизации. В отличие от простого случайного поиска, они основаны на принципах, заимствованных у природы. Это механизмы генетической наследственности и естественного отбора. Впервые ГА был предложен Гольдбергом в 1989 г. на основе идей, изложенных Дж. Холландом в своей работе «Адаптация в естественных и искусственных системах» (1975) [1].

Основная идея генетического алгоритма состоит в создании популяции особей (индивидов), каждая из которых представляется в виде хромосомы [4]. Любая хромосома есть возможное решение рассматриваемой оптимизационной задачи. Для поиска лучших решений необходимо только значение целевой функции, или функции приспособленности. Значение функции приспособленности особи показывает, насколько хорошо подходит особь, описанная данной хромосомой, для решения задачи.

Хромосома состоит из конечного числа генов, представляя генотип объекта, т.е. совокупность его наследственных признаков. Процесс эволюционного поиска ведется только на уровне генотипа. К популяции применяются основные биологические операторы: скрещивания, мутации, инверсии и др. В процессе эволюции действует известный принцип «выживает сильнейший». Популяция постоянно обновляется при помощи генерации новых особей и уничтожения старых, и каждая новая популяция становится лучше и зависит только от предыдущей. Основное отличие ГА от традиционных (точных) методов поиска оптимумов состоит в том, что ГА с каждой эпохой улучшает оптимальное решение, но не гарантирует нахождение лучшего за конечный промежуток времени.

Классическая схема ГА включает следующие шаги [1].

1. Задать начальный момент времени $t = 0$. Создать начальную популяцию из k особей (размер популяции) $P_0 = \{A_1, A_2, ..., A_k\}$.
2. Рассчитать приспособленность каждой особи $F(A_i), i = \overline{1,k}$.

3. Выбрать из популяции две особи $A_i, A_j, i \neq j$.
4. Выполнить оператор кроссовера и мутации. В результате генерируется новая особь *B* (потомок). В некоторых модификациях алгоритма после скрещивания могут быть созданы несколько потомков (например, два).
5. Поместить полученную хромосому (или несколько хромосом) в новую популяцию P_{t+1}.
6. Выполнить оператор редукции, т.е. сократить размер новой популяции до исходного размера.
7. Выполнить шаги, начиная с п. 3, *k* раз.
8. Увеличить номер текущей эпохи $t = t + 1$.
9. При срабатывании условия останова завершить работу, иначе перейти на шаг 2.

Таким образом, генетический алгоритм можно записать в виде кортежа:

$$GA =< P_0, k, n, S, C, M, F, t >,$$

где P_0 – исходная популяция; *k* – ее размер; *n* – количество генов в хромосоме; *S*, *C*, *M* – операторы отбора, кроссовера и мутации соответственно; *F* – функция приспособленности (фитнеса); *t* – критерий останова.

Для представления генотипов особей в популяции используются различные схемы кодирования. Наиболее распространенным является двоичное кодирование. Все генетические операции проводятся исключительно на уровне генотипа, т.е. с битовой строкой, а фенотип объекта используется при определении приспособленности особи.

При старте ГА начальная популяция, как правило, формируется случайным образом. Далее из популяции выбираются родители – две хромосомы, к которым будет применен оператор скрещивания (кроссовера), т.е. работает оператор отбора (селекции) *S*. Наиболее часто в ГА применяют следующие типы оператора отбора [3].

1. Вероятностный выбор особей. Случайным образом выбирают две хромосомы $A_j, A_{j+1} \in P_t$, $A_j \neq A_{j+1}$, *k* – размер популяции.

   При таком способе селекции частота образования пары зависит только от численности популяции и не зависит от функции приспособленности особи.
2. Турнирный отбор является модификацией вероятностного отбора. Случайно выбираются несколько особей из популяции (например, две) и победителем выбирается особь с наибольшей приспособленностью. Данная процедура повторяется дважды, для выбора каждого родителя.

3. Пропорциональный отбор с использованием функции фитнеса, который моделирует природный принцип «выживает сильнейший». Вероятность выбора особи для скрещивания вычисляется по формуле

$$\rho(A_i) = \frac{F(A_i)}{\sum_{j=1}^{k} F(A_j)}. \quad (1)$$

Согласно (1) вероятность передачи признаков более приспособленными особями возрастает, и одна особь может быть выбрана для нескольких родительских пар.

Оператор скрещивания (кроссовер) моделирует природный механизм передачи наследственной информации от родителей к потомкам. Он также имеет множество реализаций. В классическом операторе кроссовера Холланда результатом скрещивания двух хромосом $C_1 = (c_1^1 ... c_n^1)$ и $C_2 = (c_1^2 ... c_n^2)$ являются два потомка $H_1 = (c_1^1,...,c_i^1,c_{i+1}^2,...,c_n^2)$ и $H_2 = (c_1^2,...,c_i^2,c_{i+1}^1,...,c_n^1)$, где i – случайное натуральное число из множества $\{1,...,n-1\}$ (точка разбиения).

Оператор кроссовера выполняется с некоторой вероятностью, называемой вероятностью скрещивания. Обычно она изменяется в пределах от 0,6 до 1,0.

Оператор мутации предназначен для поддержания разнообразия особей в популяции и предотвращения преждевременного сжатия пространства поиска. Он выполняется с некоторой вероятностью (вероятностью мутации) для каждого бита хромосомы и заключается в его инвертировании. В некоторых реализациях ГА мутации может быть подвергнут только один случайно выбранный бит хромосомы. Вероятность мутации выбирается достаточно малой, обычно в диапазоне $10^{-2}…10^{-4}$.

При формировании новой популяции работает оператор отбора полученных в результате скрещивания и мутации новых индивидов. Новая популяция должна быть сокращена до исходного размера, и для этого из нее удаляются менее приспособленные особи, т.е. с наименьшими значениями функции фитнеса. На этапе отбора может использоваться стратегия элитизма, которая заключается в том, что часть особей (в простейшем случае – одна) с наибольшей приспособленностью гарантированно переходят в новую популяцию без всяких изменений [3].

Критерием останова генетического алгоритма может выступать ограничение на максимальное количество поколений, или когда средняя приспособленность популяции, определяемая по формуле

$$\overline{F}(H) = \frac{\sum_{i=1}^{k} F_i(H)}{k},$$

перестает изменяться заданное число эпох.

Высокая эффективность отыскания глобального минимума или максимума генетическим алгоритмом теоретически обоснована в фундаментальной теореме генетических алгоритмов («теореме о шаблоне»), доказанной Холландом [1]. С ее помощью доказано, что эффективность отыскания оптимума генетическим алгоритмом определяется выбранной схемой кодирования, используемыми операторам селекции, кроссовера и мутации параметрами.

В настоящее время генетические алгоритмы хорошо исследованы и широко применяются для решения комбинаторных и оптимизационных задач в экономике, технике и науке. Разработано большое количество разновидностей и модификаций простого ГА, учитавающих специфику конкретных задач: осровные, микро, ниши, клеточные, коэволюционные, иерархические, гибридные [3].

Адаптация популяционно-генетических методов к задачам повышенной размерности. Несмотря на многочисленные достоинства, все популяционно-генетические алгоритмы имеют ряд недостатков, которые мешают процессу поиска оптимального решения: потеря разнообразия в популяции и преждевременная сходимость. При решении задач повышенной размерности сюда добавляется проблема низкой точности полученного решения, которая особенно сильно проявляется при оптимизации непрерывных овражных функций [8]. Большая размерность поиска не позволяет исследовать за приемлемое время все пространство возможных решений, а, при нахождении хромосомы, значительно улучшающей фитнес-функцию, имеем высокую вероятность потери решения. Первая проблема решается распараллеливанием работы генетического алгоритма и специальными стратегиями генерации начальной популяции. Причем распараллеливание возможно в двух аспектах: организационный – формируется несколько независимых субпопуляциий, которые периодически обмениваются лучшими особями, и вычислительный – работа субпопуляций ведется на многопроцессорных вычислительных системах с массовым параллелизмом, чем достигается масштабируемость, близкая к линейной.

На практике генетические алгоритмы нередко используют совместно с другими методами, которые позволяют повысить их точность, разрабатывая так называемые *гибридные* схемы. Очень часто объединение ГА с традиционными методами дает синергетический эффект в виде высокой точности и скорости получаемых решений. Так, в работе [9]

один из авторов этой статьи исследовал гибридный генетический алгоритм для решения задач оптимизации в непрерывных пространствах. Он основан на параллельной работе генетических операторов и какого-либо градиентного метода. В популяции, созданной генетическим алгоритмом, выбирается лучшая особь – лидер. Этот лидер обучается отдельно по градиентному методу. Если его качественный показатель при этом лучше, чем у всех остальных особей в популяции, то он вводится в популяцию и участвует в воспроизводстве потомков. Если же появляется особь в популяции, полученная в результате эволюции, с лучшим показателем, то лидером становится она. Исследование предлагаемого гибридного алгоритма проводилось на основе двух видов генетических алгоритмов (в бинарных и вещественных кодах) и трех типов градиентных методов (метод наискорейшего спуска, метод сопряженных градиентов, квазиньютоновский метод или метод переменной метрики). Тестирование на сложных овражных функциях показало его высокую эффективность. За несколько сотен итераций им был найден глобальный минимум функции Розенброка размерностью 2000 с точностью 10^{-5} (известные алгоритмы оптимизации не справились с задачей такой размерности).

Исходя из вышесказанного, можно предположить о большом потенциале гибридных схем традиционных методов дискретной оптимизации и генетических алгоритмов. Такие попытки уже делались. Так, в работе [10] предложена гибридная схема ГА и метода ветвей и границ для решения задач календарного планирования. В нем вводятся несколько эвристик на начальном этапе поиска для определения верхней границы и далее используется генетический алгоритм в процессе уточнения значений верхней границы. Эксперименты авторов показывают, что оптимальность ветвей и границ в этом случае достигается чаще. Еще в одной работе [11] решается эта же задача, но с точки зрения сокращения времени поиска, которое удается сократить в среднем на 25% при помощи использования генетического алгоритма.

Таким образом, анализ текущего состояния проблемы показывает, что исследование новых эффективных методов для дискретной оптимизации в пространствах повышенной размерности должна вестись в области разработки гибридных схем с применением эволюционных алгоритмов.

Литература

1. Бершицкий Ю.И., Нечаев В.И., Горячев Ю.О., Бондаренко В.В. Расчет технико-экономических характеристик машинно-тракторного парка сельхозпредприятий. РОСПАТЕНТ. Свидетельство № 2006611933.

2. Holland J.H. Adaptation in Natural and Artificial Systems. – The University of Michigan Press, University of Michigan, Ann Arbor, 1975.
3. Скобцов Ю.А. Основы эволюционных вычислений. Учебное пособие. – Донецк: ДонНИУ, 2008.
4. Подлазова А.В. Генетические алгоритмы на примерах решения задач раскроя // Информационные технологии в управлении. – № 2. – 2008. – С. 57-63.
5. Неймарк Е.А. Адаптация оптимальных решений нестационарных комбинаторных задач с помощью популяционно-генетических методов // Автореф. дис. на соиск. уч. степ. канд. техн. наук; НГУ им. Лобачевского. – Н. Новгород, 2008.
6. Кочетов Ю.А. Методы локального поиска для дискретных задач размещения // Автореф. дис. на соиск. уч. степ. канд. физ.-мат. наук; Ин-т математики им. С.Л. Соболева СО РАН. – Новосибирск, 2009.
7. Нгуен М.Х. Разработка и реализация численных методов решения оптимизационных задач большой размерности // Автореф. дис. на соиск. уч. степ. канд. физ.-мат. наук; Вычислительный центр им. А.А. Дородницына РАН. – Москва, 2009.
8. Сергиенко И. В. Математические модели и методы решения задач дискретной оптимизации. – Киев: Наукова Думка, 1988.
9. Паклин Н.Б., Сенилов М.А., Тененев В.А. Интеллектуальные модели на основе гибридного генетического алгоритма с градиентным обучением лидера // Искусственный интеллект. – Донецк: Наука і освіта. – 2004. – № 4. – С. 159-168.
10. Portmann M.-C., Vignier A. Branch And Bound Crossed With GA To Solve Hybrid Flowshops // European Journal of Operational Research. – 1998. – V. 107. – №2. – P. 389-400.
11. Morita H., Shio N. Hybrid Branch and Bound Method with Genetic Algorithm for Flexible Flowshop Scheduling Problem // JSME International Journal Series C. – Vol. 48 (2005). – №1. – P.46-52.

ИНТЕЛЛЕКТУАЛЬНЫЙ АНАЛИЗ ДАННЫХ В СИСТЕМЕ ПРОТИВОДЕЙСТВИЯ РАСПРОСТРАНЕНИЮ ЭПИДЕМИЙ ГРИППА

***Боев Б.В.**, зав. лаб. Эпидемиологической кибернетики НИИЭМ им. Н.Ф. Гамалеи, **Болотова Л.С.,** профессор, **Демина Н.Н.**, студент, Российский государственный университет инновационных технологий и предпринимательства, г. Москва*

Среди масштабных бед и катастроф, сопровождающих всю историю человечества, наряду с голодом, войнами и стихийными бедствиями, первостепенное значение имеют эпидемии тяжелых инфекционных болезней. По своей социальной значимости, огромному ущербу, наносимому здоровью населения и экономике, грипп находится на первом месте среди всех заболеваний человека [1]. Заболеваемость гриппом и гриппоподобными инфекциями превышает суммарную заболеваемость всеми остальными инфекциями. Нет ни одной болезни человека, сравнимой в настоящее время по этому показателю с гриппом. На долю гриппа и ОРЗ приходится 10-30 % временной нетрудоспособности населения. В отдельные годы грипп и ОРЗ составляли до 40 % всех заболеваний взрослых, зарегистрированных в поликлинике, более 80 % всей инфекционной патологии, более 60 % заболеваний среди детей. Наибольшая смертность от гриппа регистрируется в период эпидемий, вызываемых новыми антигенными вариантами вируса гриппа А. Эпидемии гриппа существенно влияют на общую смертность населения и смертность от соматических заболеваний. Установлено, что эпидемии гриппа сопровождаются выраженной дополнительной смертностью населения от других заболеваний. Помимо вреда, наносимого непосредственно здоровью населения, эпидемии гриппа дают значительный экономический ущерб экономике стран за счет потерь рабочего времени, выплат по социальному страхованию заболевшим и затрат на лечебные и профилактические мероприятия.

В СССР в НИИЭМ им Н.Ф. Гамалеи РАМН ещё в 60-е годы было организовано научное направление исследований, связанное с оперативным анализом и прогнозом массовых эпидемий на основе компьютерного моделирования. В институте была создана лаборатория эпидемиологической кибернетики, которая была и остается ведущим центром России по разработке и применению математических и компьютерных моделей для изучения эпидемий, анализа и прогноза их социально-

экономических последствий [9]. Компьютерный инструментарий прогнозирования эпидемий гриппа, используемый в лаборатории, основан на модели эпидемии гриппа типа Кермака-МакКендрика с феноменологией инфекции типа SEIRF [3]. В работе использовались соотношения математической модели эпидемии гриппа [2], которая определена 5-ю стадиями-состояниями развития гриппозной инфекции (схема инфекции **S→E→I→[R,F]**), где: **S** – восприимчивые к инфекции люди, **E** – зараженные люди в инкубации, **I** – инфекционные, заразные больные, **R** – переболевшие, иммунные (реконвалесценты), **F** – умершие от осложнений больные.

В этой модели все расчеты по анализу и прогнозу эпидемического процесса гриппа выполняются по системе нелинейных дифференциальных уравнений с соответствующими начальными условиями эпидемии. Данная модель обеспечивает хорошую точность анализа процессов реальной эпидемии гриппа (верификация модели) и прогнозирования эпидемического процесса гриппа при различных сценариях реализации мер противодействия – программа вакцинации населения, а также экстренной профилактики, диагностики, терапии, изоляции инфекционных больных, мер по социальному воздействию на группы риска населения. В настоящее время ввод значений параметров эпидемии (доля восприимчивого населения, интенсивность взаимодействия между людьми) и перебор вариантов противодействия распространению эпидемии в программе SEIRF производится вручную – специалистом-эпидемиологом (экспертом) на основании своих знаний и опыта. Это обстоятельство вносит субъективизм в выбор значений параметров моделирования, делает невозможным исследование влияния на развитие эпидемии всех возможных сочетаний мер противодействия, затрудняет обоснование необходимости применения тех или иных мер, полученных по итогам моделирования.

Для устранения этих проблем в 2009 мы начали разработку системы противодействия распространению эпидемий гриппа (СПЭГ), взяв за основу пассажиропоток воздушного транспорта. Структура СПЭГ представлена на рис. 1. В ней выделены следующие подсистемы:

- модуль сбора и анализа информации о воздушных полетах;
- база данных городов мира с характеристиками;
- математическая модель эпидемии гриппа;
- база знаний с правилами оценки эпидемиологической опасности территории (города, района и т.п.);
- система поддержки принятия решений и выдачи рекомендаций по выбору мер противодействия эпидемии.

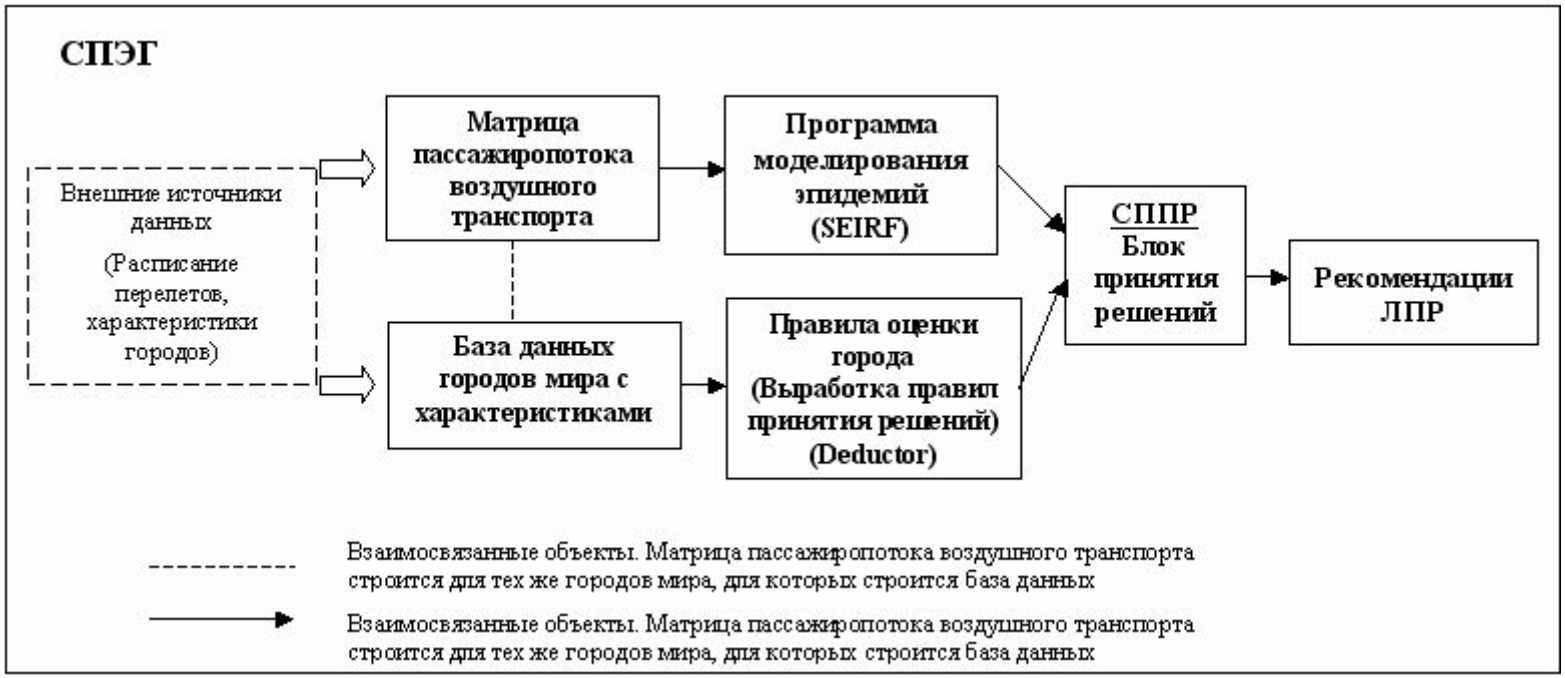


Рис. 1. Структура СПЭГ

Рассмотрим кратко основную схему функционирования СПЭГ.

Из внешних источников данных (сайт международного расписания авиаперелетов, сайты с информацией по городам мира и т.д.) извлекаются сведения о международных авиаперелетах и городах мира, для которых проводиться моделирование и которые используются в подсистеме анализа трафика пассажиров воздушного транспорта для составления базы данных городов мира с характеристиками городов. В подсистеме анализа трафика пассажиров воздушного транспорта из множества входных источников, содержащих информацию по перелетам между городами, формируется матрица полетов за день (месяц, год). Матрица поступает в программу моделирования эпидемий SEIRF, где с помощью нее прогнозируется распространение эпидемии из города-очага заболеваемости по миру. Отдельно, в специально созданном хранилище данных на базе аналитической платформы *Deductor* и ряда БД с характеристиками территорий, формируется многомерная БД с информацией о численности населения городов, экологии и других, важных для эпидемиологической оценки параметров. Эта БД анализируется с применением методов ИАД [4]. В результате выявляются неявные скрытые закономерности в данных, на основе которых строятся правила оценки эпидемиологической опасности городов.

Данные по численности населения городов мира были взяты с крупнейшего портала по мировой статистике населения городов – *World Gazetteer* [5].

Климат городов мира определялся по классификации профессора МГУ Б.П.Алисова, согласно которой на Земле существует 7 типов климатов [6], составляющих климатические пояса. Четыре из них являются основными, а три – переходными. К основным типам относятся: экваториальный; тропический; умеренный; полярный. К переходным относятся: субэкваториальный; субтропический; субполярный. В контексте рассматриваемых городов и для сужения пространства анализируемых признаков количество учитываемых климатических видов снижено до четырех: тропический, умеренный, субтропический, континентальный.

Характеристика городов по уровню жизни населения, используемая в базе данных городов мира, основывалась на статистике уровня жизни той страны, к которой относиться данный город. Оценка уровня жизни страны, взята из «Доклада о развитии человека 2009» подготовленном Организацией Объединенных Наций (ООН) [7]. В Докладе о развитии человека традиционно содержится информация об индексе развития человека, который ежегодно рассчитывается экспертами Программы развития ООН (ПРООН) совместно с группой независимых международных экспертов. Индекс измеряет достижения страны с точки зрения продол-

жительности жизни, получения образования и фактического дохода. Достойный уровень жизни измеряется величиной валового внутреннего продукта (ВВП) на душу населения в долларах США по паритету покупательной способности (ППС). Эти измерения стандартизируются в виде числовых значений от 0 до 1, среднее арифметическое которых представляет собой совокупный показатель HDI в диапазоне от 0 до 1. Затем страны ранжируются на основе этого показателя, и первое место в рейтинге соответствует наивысшему значению HDI. В результате по уровню жизни страны делятся на 4 класса страны с уровнем жизни: очень высоким, высоким, средним, низким.

Характеристика городов по уровню экологии, основывается на рейтинге, составленном по экологическому индексу *Evironmental Performance Index* за 2010 год [8]. Такую оценку стран мира по экологическому показателю разработала группа американских ученых на базе Йельского и Колумбийского университетов. *Environmental Performance Index (EPI)* – это общий для всех стран экологический индекс, разработанный для того, чтобы измерить состояние окружающей среды по общим для всех критериям. По данному рейтингу страны подразделяются на страны со следующими уровнями экологии: очень высоким – 1; высоким – 2; средним – 3; низким и очень низким – 4.

Многомерная БД построена на основе хранилища данных *Deductor Warehouse* аналитической платформы *Deductor* на базе СУБД *Firebird 1.5* [4]. Структура хранилища представлена на рис. 2.

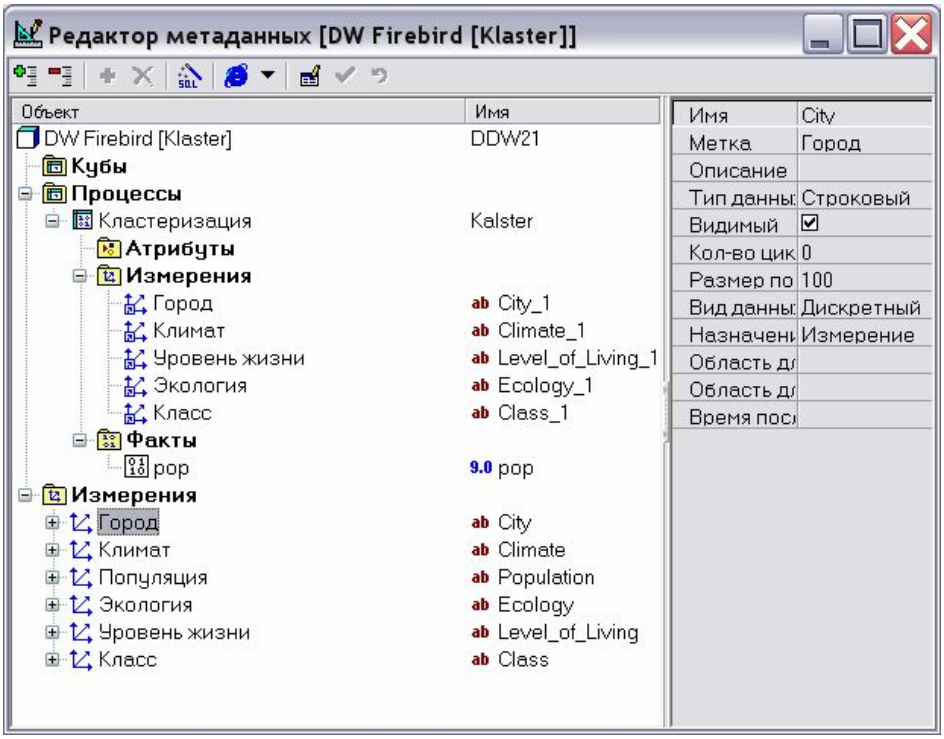


Рис. 2. Структура хранилища данных

Данная структура содержит 5 измерений: *Город*, *Климат*, *Экология*, *Уровень жизни*, *Популяция (pop)*, *Классы (1-4)* и соответствующие ссылки на них. Данные о городах мира и все остальные данные загружаются в хранилище из текстового файла, который был сформирован ранее.

После загрузки данных в хранилище, используя *визуализатор Таблица*, получаем БД по 155 городам мира.

На следующем шаге к этой БД применяются алгоритмы анализа для выявления правил оценки эпидемиологической опасности городов.

Так, на основании данных о заболеваемости по 105 городам из 155 изучаемых (из разных регионов земли, Европы, Америки, Азии и т.д.) все города мира были разбиты на классы с уровнями эпидемиологической опасности: низкий –1; средний –2; высокий –3; чрезвычайно высокий – 4.

Эти 105 городов послужили обучающей и тестовой выборкой для алгоритма кластеризации «Самоорганизующаяся Карта Кохонена» [5].

В результате определились 4 кластера городов с перечисленными уровнями эпидемиологической опасности.

После запуска метода дерева решений на БД городов мира были получены следующие правила, представленные ниже в виде дерева решений (рис. 3).

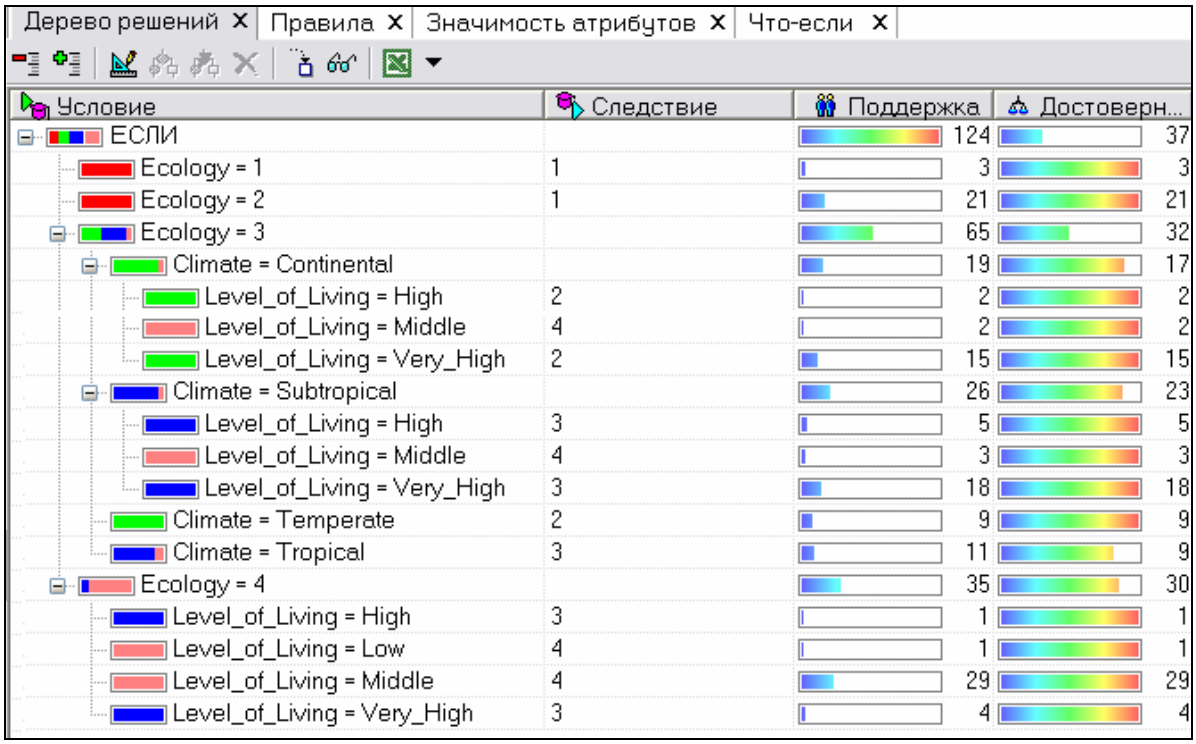


Рис. 3. Дерево решений

Анализируя полученное дерево можно сказать, что все классы получились чистыми, то есть не содержат примесей (объектов других классов). В процессе построения дерева при определении параметра ветвления на каждом уровне иерархии, по которому происходит разделение на дочерние узлы, мы использовали критерий наибольшего устранения неопределенности. Более значимые факторы, по которым проводится классификация, находятся на более близком расстоянии (глубине) от корня дерева, чем менее значимые. В нашем случае, наиболее значимым фактором (наиболее эффективным для ветвления) является параметр *Экология*, затем фактор *Климат*. А фактор *Уровень жизни* значим только в сочетании с другими факторами. Еще один интересный вывод – это отсутствие в построенном дереве параметра *Популяция*. Видимо, это потому, что все 155 городов имеют разное количество населения, и алгоритм не смог вывести критерий оценки для параметра популяции, поэтому он не повлиял на оценку эпидемиологической опасности города. Приведём примеры правил:

1. ЕСЛИ((*Экология* = 3) И (*Климат* = Континентальный) И (*Уровень жизни* = Высокий)) ТО (*Класс* = 2 (эпидемиологическая опасность средняя))
2. ((ЕСЛИ *Экология* = 4) И (*Уровень жизни* = Средний) ТО (*Класс* = 4 (эпидемиологическая опасность чрезвычайно высокая))
3. ЕСЛИ ((*Экология* = 3) И (*Климат* = Тропический) ТО (*Класс* = 3 (эпидемиологическая опасность высокая)).

Дополнительно правила были подтверждены и дополнены методом ассоциаций. Таким образом, правила установления соответствия были подтверждены разными методами анализа, что с уверенностью позволяет говорить о том, что они имеют право быть включенными в БЗ для СППР (рис. 4).

На следующем шаге специальная программа – информационная система анализа трафика пассажиров воздушного транспорта – формирует матрицу пассажиропотока между 155 городами мира. Для этого была разработана программа анализа трафика пассажиров воздушного транспорта, работающая в двух режимах: создания общего файла расписания городов и последующего формирования матрицы пассажиропотока, либо формирования матрицы пассажиропотока из уже готового общего файла расписания (с информацией по всем исследуемым городам).

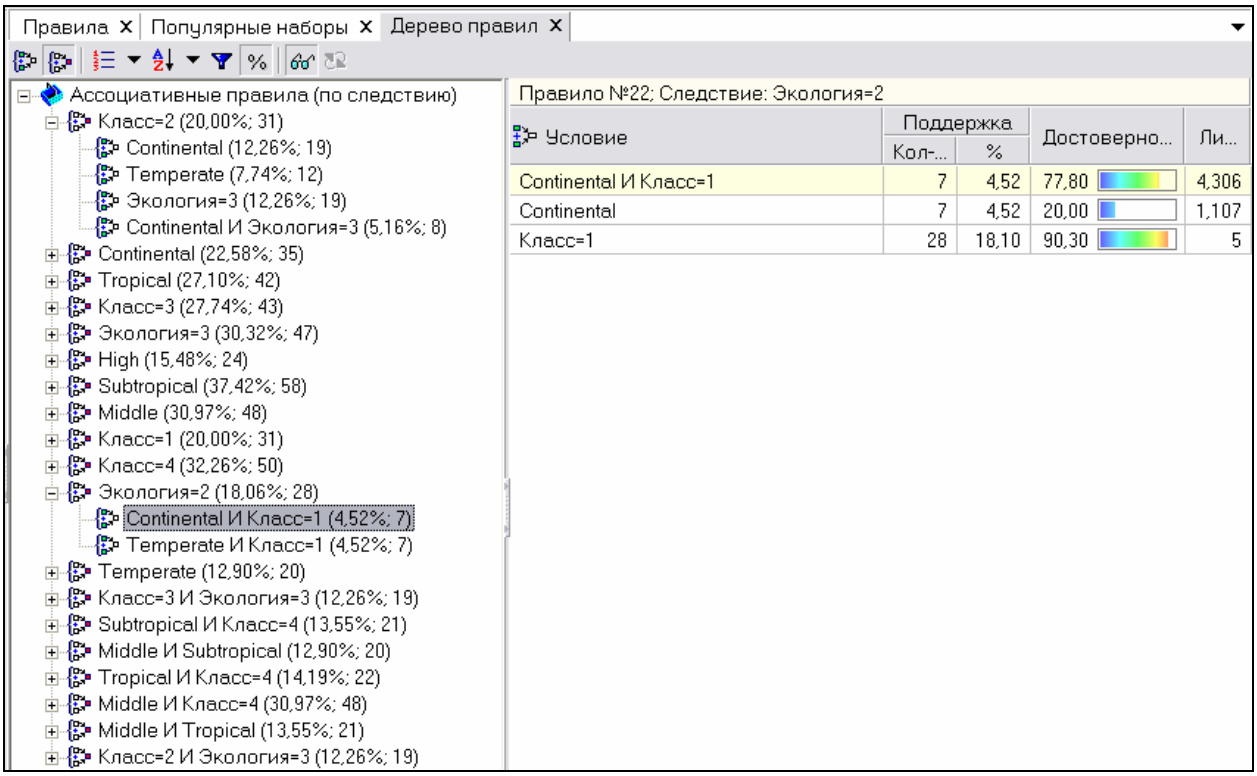


Рис. 4. Ассоциативные правила

Формат матрицы строго согласован со структурой интерфейса программы моделирования эпидемий SEIRF, которая работает на следующем шаге. Проведенный анализ правил и экспертная оценка эпидемиологической опасности городов показали, что полученные правила верны и отражают реальную картину заболеваемости населения в городе. Например, в класс 1 с низкой эпидемиологической опасностью попали самые благополучные, экономически развитые, с отличным уровнем медицины такие города, как: Цюрих, Женева, Гамбург, Лондон, Париж, Вена. В класс 2 – средняя эпидемиологическая опасность попали такие города, как: Москва, Санкт-Петербург, Минск. В класс 3 с высокой эпидемиологической опасностью попали промышленные города США и прочие «средние» города. В класс 4 с чрезвычайно высокой эпидемиологической опасностью попали города Азии и Африки, с чрезвычайно низким общим уровнем жизни и высокой заболеваемостью, низким уровнем медицины.

Для моделирования эпидемии с использованием программы SEIRF необходимо назначить городу, в котором происходит зарождение эпидемии, коэффициенты иммунитета (доля восприимчивого населения) и коммуникабельности (интенсивность взаимодействия).

В соответствии с полученными правилами и на основании опыта экспертов-эпидемиологов классам городов, а, следовательно, и городам,

которые принадлежат этим классам, были назначены следующие коэффициенты:

1 класс – город с низкой эпидемиологической опасностью; доля восприимчивых – 0,5; интенсивность взаимодействия – 0,5.

2 класс – город со средней эпидемиологической опасностью; доля восприимчивых – 0,61; интенсивность взаимодействия – 0,7.

3 класс – город с высокой эпидемиологической опасностью; доля восприимчивых – 0,85; интенсивность взаимодействия – 0,8.

4 класс – город с чрезвычайно высокой эпидемиологической опасностью; доля восприимчивых – 0,91; интенсивность взаимодействия – 0,9.

Результаты моделирования показаны ниже. На рис. 5 представлен внешний вид карты до начала эпидемии. Для примера в качестве города-«источника» эпидемии был задан Мехико. Доля восприимчивых, исходя из правил оценки эпидемиологической опасности города Мехико, задана равной 0,61, интенсивность взаимодействия – 0,9. На рис. 6 представлен прогноз развития эпидемии и роста числа заболевших по 155 города мира, а на рис. 7 – ее начало.

Далее эпидемия распространяется, и заболевание постепенно разносится по оставшимся городам. В городах, где пик заболевания уже пройден, эпидемия постепенно затухает (рис. 8).

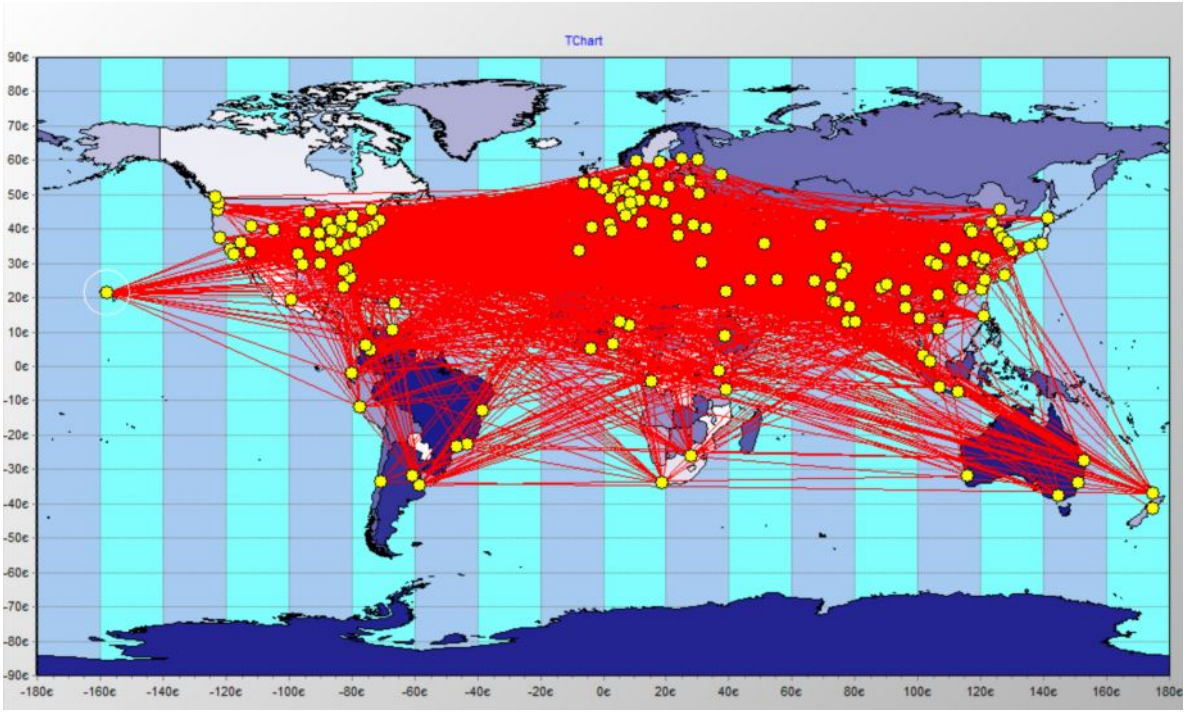


Рис. 5. Эпидемия еще не началась

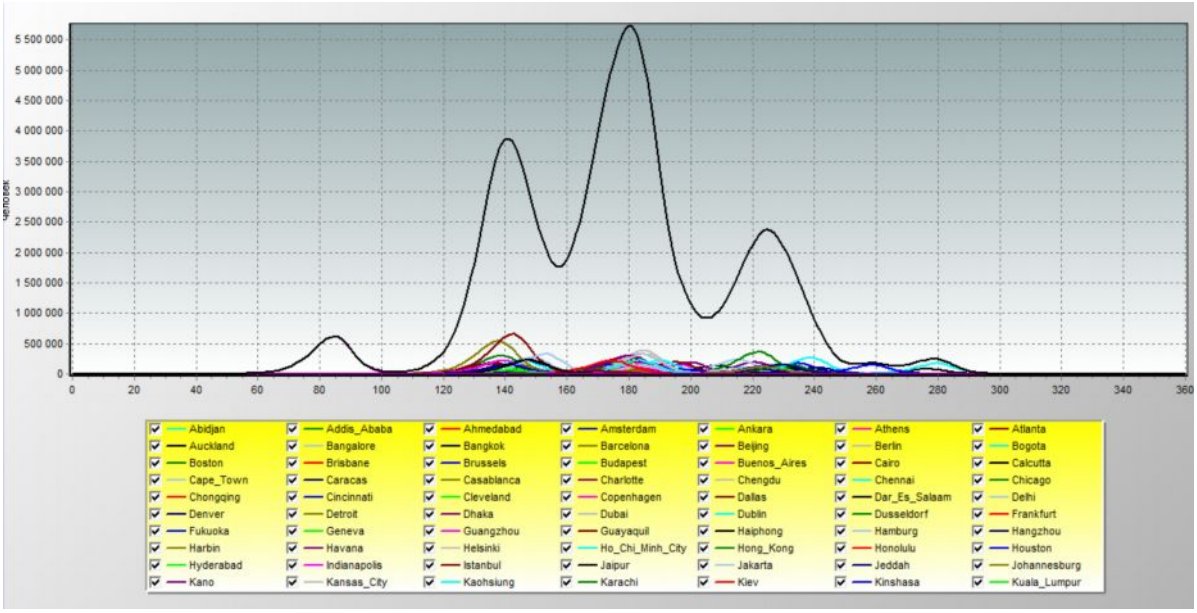


Рис. 6. Прогноз развития

На следующем этапе предполагается реализовать подсистему СПЭГ – СППР, которая будет предлагать наиболее эффективные варианты противодействия развитию эпидемии, исходя из конкретных условий того региона, в котором всё это происходит. При этом предполагается учитывать экономические возможности региона, наличие запасов медикаментов и их размещение (скорость доставки), возможности медицинского персонала и больниц и другие экономические и социальные параметры. В качестве модели представления знаний и аппарата реализации предполагается использовать возможности нечётких систем.

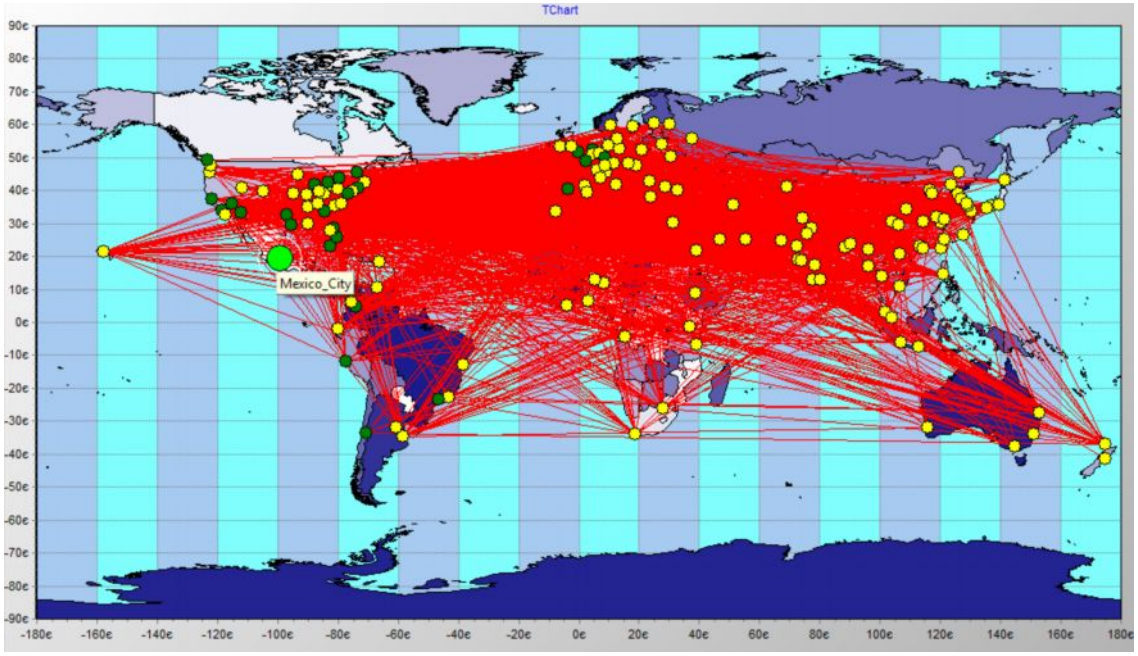


Рис. 7. Начало распространения эпидемии

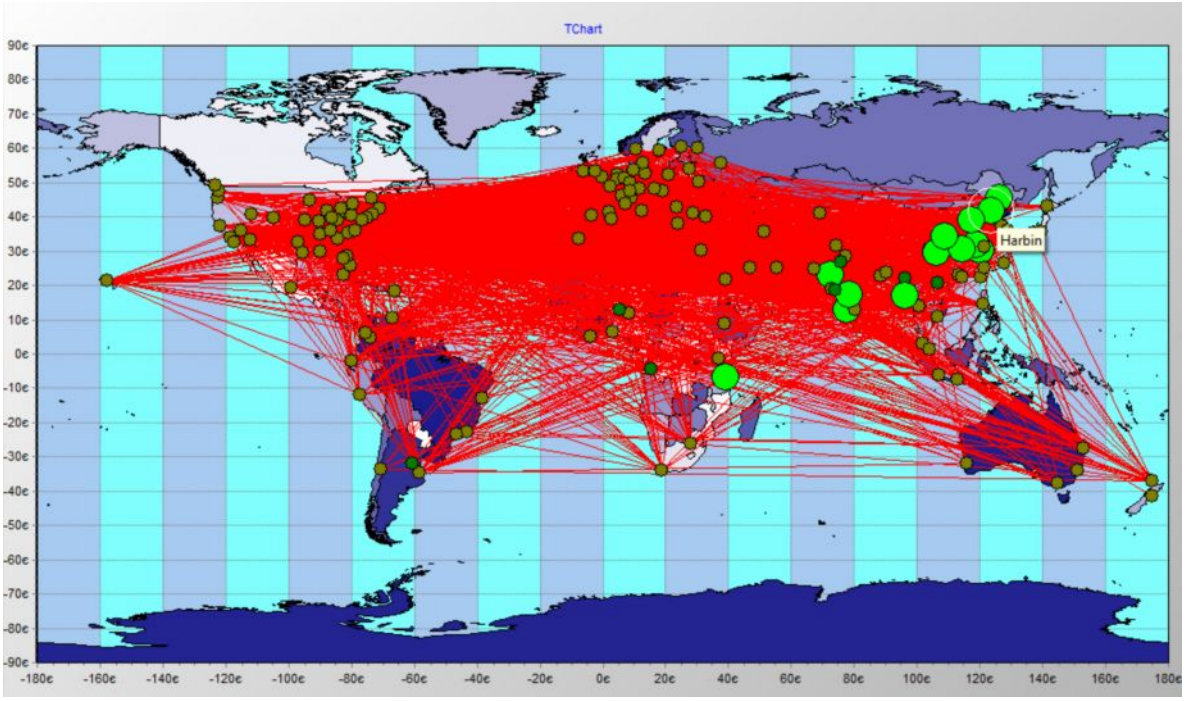


Рис. 8. Дальнейшее развитие эпидемии

Литература

1. Смородинцев А.А. Грипп и его профилактика. – Л.: Медицина, 1984. – 383 с.
2. Федеральный конституционный закон «О чрезвычайном положении» от 30 мая 2001 г.
3. Боев Б.В. Современный этап математического моделирования в эпидемиологии инфекционных заболеваний // Эпидемиологическая кибернетика: модели, информация, эксперименты: Сборник науч. трудов. – М., 2010. – С. 6
4. Паклин Н.Б., Орешков В.И. Бизнес-аналитика: от данных к знаниям. – СПб.: Питер, 2009. – 624 с.
5. Статистический портал World Gazetteer [Электронный ресурс]/ World Gazetteer . – 2010 г. Режим доступа: www.world-gazetteer.com, свободный. – Яз. Рус.
6. Максаковский В.П. Географическая картина мира. – М.: Дрофа, 2009. – 480 с.
7. Доклад о развитии человека 2009 [Электронный ресурс]/ ООН . – 2010 г. Режим доступа: http://www.un.org/ru/development/hdr/2009/, свободный. – Яз. Англ.
8. Country scores [Электронный ресурс] / Environmental Performance Index. – 2010 г. Режим доступа: http://epi.yale.edu/Countries, свободный. – Яз. Англ.
9. http://www.gamaleya.ru.

ПРИМЕНЕНИЕ СОВРЕМЕННЫХ ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ И ИНТЕЛЛЕКТУАЛЬНЫХ МЕТОДОВ АНАЛИЗА В ЗАДАЧЕ ОЦЕНКИ НЕДВИЖИМОСТИ

***Медведева Т.В.**, студент, **Прокопенко Н.Ю.**, доцент, Нижегородский Государственный архитектурно-строительный Университет, г. Нижний Новгород*

В настоящее время среди элементов рыночной экономики недвижимость занимает особое место, выступая как в качестве средств производства, так и в качестве объекта потребления. Залог успеха при принятии управленческих решений государственными органами управления и риэлтерскими компаниями – правильное определение рыночной стоимости объектов недвижимого имущества.

Рыночная стоимость объекта недвижимости определяется при помощи трех подходов: 1) подход с точки зрения затрат; 2) оценка по прямому сравнению продаж; 3) подход с точки зрения доходности. Перечисленные подходы реализуются с помощью различных экспертных методов, а также методов математической статистики, программно реализованных методов кластерного, регрессионного, факторного анализа и многих других. В последнее время становятся популярными методы интеллектуального анализа или методы Data Mining.

Data Mining – это процесс обнаружения в данных ранее неизвестных, нетривиальных, практически полезных и доступных интерпретации знаний, необходимых для принятия решений в различных сферах человеческой деятельности. Информация, найденная в процессе применения методов Data Mining, является нетривиальной и ранее неизвестной. Знания должны описывать новые связи между свойствами, предсказывать значения одних признаков на основе других.

В данной статье описываются современные подходы к решению задач оценки недвижимости вторичного рынка г. Нижнего Новгорода.

При анализе ценовой ситуации на рынке решаются следующие задачи:

- анализ состояния Нижегородского рынка недвижимости;
- исследование факторов, влияющих на изменения стоимости;
- прогнозирование тенденций изменения цен.

В качестве информационной базы используются данные веб-сайта агентства недвижимости «Волгожилстрой» и федерального портала о недвижимости «Мир квартир» о продаже однокомнатных квартир вторич-

ного рынка недвижимости одного из районов г. Нижнего Новгорода, а также материалы печатного и электронного вариантов газеты «Из рук в руки».

За последние 10 лет (январь 2000 г. – июнь 2010 г.) динамика стоимости вторичного жилья г. Нижнего Новгорода менялась в зависимости от экономических изменений в целом по стране и в мире. В период с 2000 г. по 2005 г. наблюдался стабильный рост цен на недвижимость, который объясняется общим приростом благосостоянием населения России, а с июля 2008 г по июнь 2009 г. наблюдался спад цен на недвижимость из-за последствий мирового финансового кризисам. С июня 2009 по июнь 2010 года наблюдалось замедление темпов падения цен (рис. 1).

Средняя стоимость квадратного метра общей площади жилья зависит не только от экономической ситуации по стране и в мире, а также от таких критериев как районность, комнатность, тип планировки и материал цен. Ограниченность предложения, с одной стороны, а также повышение степени недоверия к строительным компаниям – с другой («замороженные стройки» и др.), приводят покупателя на вторичный рынок жилья.

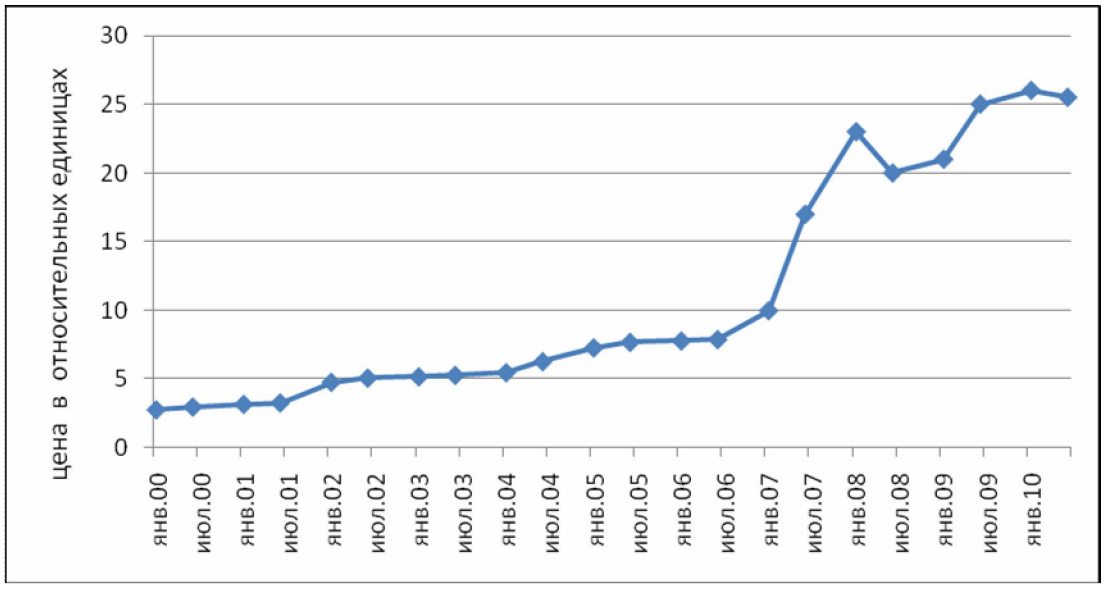


Рис. 1. Динамика стоимости вторичного жилья (руб./кв.м.) в Нижнем Новгороде за период январь 2000 г. – июнь 2010 г.

С точки зрения оценки недвижимости актуальными являются две основные задачи – классификация объектов недвижимости и оценка стоимости жилой недвижимости.

В качестве инструментального средства поддержки принятия решений в области оценки недвижимости на основе интеллектуальных

средств была выбрана аналитическая платформа *Deductor Academic*, обеспечивающая большой набор средств для аналитической обработки, статистического анализа, манипулирования, визуализации данных, кластеризации, прогнозирования и многих других технологий интеллектуального анализа данных.

Анализ данных в аналитической платформе *Deductor* базируется на построении сценариев обработки. Используя аналитическую платформу, был создан сценарий оценки недвижимости, включающий модели классификации и прогнозирования цены на основе множественной линейной регрессии, деревьев решений и нейронных сетей.

Сценарий представляет собой иерархическую последовательность этапов обработки и визуализации наборов данных. Разработанный сценарий включает следующие шаги:

1. Импорт данных.
   - Учитывая ограниченные возможности работы с информацией версии *Deductor Academic*, данные о продажах квартир загружаются из текстового файла.
   - При загрузке информации необходимо правильно установить соответствия между входными данными (факторами, влияющими на стоимость жилья) и значениями по умолчанию.
2. Проверка качества данных.
   - При помощи визуализатора *Статистика* и различных диаграмм проверяем качество данных, и если оно не является приемлемым, переходим к следующему шагу.
3. Очистка данных.
   - С помощью обработчика данных *Дубликаты и противоречия* в исходной выборке данных выявляем дублирующие (сделки с одинаковыми параметрами квартиры и одинаковой ценой) и противоречивые записи (сделки с одинаковыми параметрами, но разными ценами).
   - При помощи визуализатора *Диаграмма* выявляем аномальные значения (информация о сделках с аномально высокой или аномально низкой ценой – следствие ошибок ввода и других случайных факторов, а также мошеннических действий, например с целью завышения страховой стоимости квартир). Для этого вначале с помощью обработчика *Калькулятор* вычисляем цену за 1 кв.м. и строим график этого показателя.
   - Проводим визуальный анализ и выявляем крайние точки, где график себя ведет неоднородно.
   - Исключаем эксцессы с помощью обработчика *Фильтрация*.

4. Построение модели классификации, используя популярный инструмент Data Mining – деревья решений.

Алгоритм конструирования дерева решений не требует от пользователя выбора входных атрибутов (независимых переменных). На вход алгоритма можно подавать все существующие атрибуты, алгоритм сам выберет наиболее значимые среди них, и только они будут использованы для построения дерева. Классификационная модель, представленная в виде дерева решений, упрощает понимание решаемой задачи, так как позволяет легко интерпретировать результаты.

- Используем для решения задачи классификации обработчик «Дерево решений». На вход дерева решений подаем характеристики квартиры, а выходом будут служить 3 класса: дорогие, средние, дешевые квартиры.
- Разбиваем все данные на три категории, в зависимости от цены за 1 кв.м. при помощи обработчика данных *Квантование*, так как в обработчике *Дерево решений* выходное значение должно быть дискретным.
- Выполняем при помощи обработчика *Замена значений* замену значений по таблице подстановок, которая содержит пары, состоящие из исходного значения и выходного значения. Например, <Цена за 1 кв.м. до 60000 >– <Дешевая квартира>, <Цена за 1 кв.м. от 60000 до 62857,15 > – <Средняя квартира>, <Цена за 1 кв.м. от 62857,15>– <Дорогая квартира>.
- Определяем класс, к которому относиться любая квартира: *Дешевая квартира*, *Средняя квартира*, *Дорогая квартира* на основании полученных обработчиком *Дерево решений* списка иерархических правил вида «Если.. , то». Например, для того, чтобы определить тип квартиры, можно рассмотреть одно из полученных правил: если одновременно: *Остальная площадь* меньше 20,5 м^2, *Жилая площадь* не менее 14 м^2 и *Кухня* не менее 5,4 м^2, то квартира, имеющая такие параметры, будет отнесена к классу средних квартир (табл. 1).

Поддержка показывает, какой процент сделок из имеющегося набора данных удовлетворяет условию правила, а достоверность показывает, какой процент сделок, удовлетворяющих условию правила, удовлетворяет и его следствию (с какой вероятностью оцениваемая квартира, удовлетворяющая условию правила, удовлетворяет и его следствию).

Таблица 1

Пример работы правила дерева решений

№	Условие			Следствие	Поддержка		Достоверность	
	Показатель	Знак	Значение	Класс	Кол-во	%	Кол-во	%
1	Остальная площадь	<	20.5	Средняя	50	62,50	26	52,00
2	Жилая площадь	>=	14					
3	Кухня	>=	5,4					

В сценарии было построено несколько моделей классификации на основе Дерева решений с разным порядком уровня доверия. Используя визуализатор Диаграмма рассеяния, было выявлено, что большая точность классификации достигается при уровне доверия 20%. С целью получения более компактного дерева, для более простой интерпретации полученных правил, уровень доверия был снижен до 5% (при этом качество этой модели ухудшилось).

5. Построение модели классификации на основе модели нейронной сети.

Нейронные сети – это класс аналитических методов, построенных на принципах функционирования мозга и позволяющих прогнозировать значения некоторых переменных после прохождения этапа так называемого обучения на имеющихся данных. Работа нейронной сети аналогична работе эксперта, который может оценить класс или стоимость объекта недвижимости только на основе его свойств (признаков).

Нейронные сети – эффективный инструмент для решения задач классификации и предсказания, если задача хорошо формализована. Задача классификации и оценки стоимости недвижимости является хорошо формализованной, так как входные данные (определенный набор стандартных признаков квартир) и желаемый выход (класс или цена) хорошо интерпретируются, а также имеется богатый опыт в виде предыдущих продаж.

- Используем для решения задачи классификации в *Deductor* обработчик *Нейросеть*. Для решения задачи классификации выбираем следующую архитектуру нейронной сети: входной слой состоит из 9 нейронов, на которые подаются значения девяти известных факторов (*Жилая площадь*, *Кухня*, *Стены*, *Этаж*, *Балкон*, *Санузел*, *Телефон*, *Состояние*, *Остальная площадь*), один скрытый слой состоит из двух нейронов, и выходной слой состоит из трех нейронов, соответствующих трем классам (*Дешевая*, *Средняя*, *Дорогая квартира*).

- Разбиваем все данные на три категории, в зависимости от цены за 1 кв.м. при помощи обработчика данных *Квантование* и выполняем при помощи обработчика *Замена значений* замену значений по таблице подстановок.
- Запускаем процесс обучения нейронной сети.
- После обучения нейронной сети проверяем при помощи визуализатора *Что-если*, как работает построенный нейросетевой классификатор. А именно, с помощью данного визуализатора покупатель или риелтор, определяет класс, к которому относиться квартира с заданными параметрами. Например, если клиент задает следующие параметры: *Жилая площадь* – 13 м^2, *Кухня* – 5 м^2, *Стены* – кирпичные, *Этаж* – 9, *Балкон* – имеется, *Санузел* – разделенный, *Телефон* – имеется, *Состояние квартиры* – отличное, то на выходе определяем класс – *Дешевая квартира*.

6. Построение модели прогнозирования на основе модели множественной регрессии.

Задача линейной регрессии заключается в нахождении коэффициентов уравнения линейной регрессии, которое имеет вид:

$$y = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_n x_n ,$$

где y – выходная (зависимая) переменная модели; x_1, x_2,…, x_n – входные (независимые) переменные; b_i – коэффициенты линейной регрессии.

В задаче прогнозирования входными переменными модели x_i являются наблюдения из прошлого, а y – прогнозируемое значение.

- С помощью обработчика «Корреляционный анализ» проводим корреляционный анализ для того, чтобы выявить и устранить мультиколлиниарность факторов, оценить зависимость выходного поля от входных факторов и устранить незначащие факторы. Принцип корреляционного анализа состоит в поиске таких значений, которые в наименьшей степени взаимосвязаны с выходным результатом. Такие факторы исключаются из результирующего набора данных практически без потери полезной информации. Критерием принятия решения об исключении является порог значимости. Если корреляция (степень взаимозависимости) между входным и выходным факторами меньше порога значимости, то соответствующий фактор отбрасывается как незначащий.
- Строим при помощи специального визуализатора матрицу корреляции.

- Устраняем незначащие факторы и повторно строим матрицу корреляции.
- Выявляем значимость атрибутов, которые в большей степени определяют стоимость квартиры при помощи визуализатора *Значимость атрибутов*. Устанавливаем, что наибольшее влияние на цену оказывают три атрибута: *Жилая Площадь* – 29 %, *Кухня* – 28 % и *Остальная Площадь* – 22 %.
- Строим модель множественной линейной регрессии для определения стоимости квартиры на основании отобранных факторов.
- Получаем коэффициенты линейной регрессии при помощи обработчика данных *Линейная регрессия*.
- Проводим тестирование модели на новых данных, а также сравнение спрогнозированной цены и фактической с помощью построения графика (рис. 2).
-

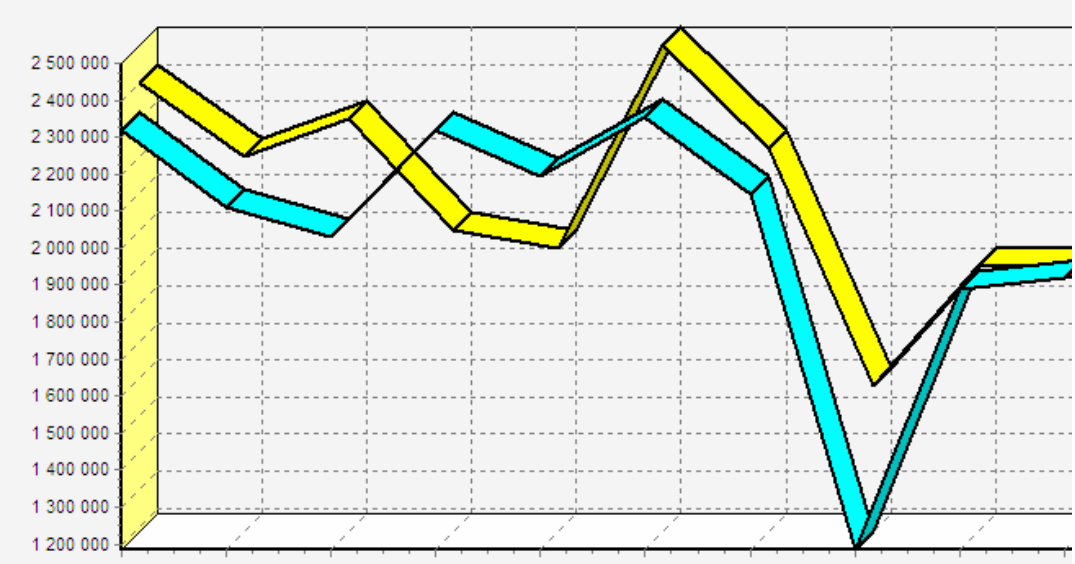


Рис.2. Диаграмма прогноза

7. Построение модели оценки стоимости жилья на основе модели нейронной сети.

- Выбираем архитектуру нейросети: входной и скрытый слой такие же, как в нейросетевом классификаторе, выходной слой имеет один нейрон, где будем получать прогнозируемое значение цены.
- Проводим обучение при помощи обработчика *Нейросеть*.
- Определяем прогнозируемую цену при помощи визуализатора *Что-если*. Например, если клиент задает желаемые параметры, то на выходе получает прогнозируемую стоимость квартиры.

8. Сравнение моделей.

- Определяем качество модели *Дерево решений* и модели *Нейронная сеть* в задаче классификации недвижимости, используя специальный визуализатор *Таблица сопряженности*, который позволяет наглядно оценить результаты классификации. Так как процент правильно распознанных примеров модели, построенной на основе нейронной сети, выше процента правильно распознанных примеров модели, построенной на основе дерева решений, делаем вывод, что при решении задачи классификации целесообразно использовать модель нейронной сети, так как она дает меньшую ошибку.
- Оцениваем качество моделей для решения задачи прогнозирования и делаем вывод, что полученные модели сравнимы по точности, так как имеют процент правильно распознанных примеров 79 % и 74 % соответственно.

Построенный сценарий позволяет автоматизировать процесс классификации и оценки стоимости объектов недвижимости. Он может быть использован как инструментальное средство в процессе принятия управленческих решений российской оценочной практики. Таким образом, благодаря современным информационным технологиям и интеллектуальным методам анализа можно облегчить и улучшить работу агентов-риэлторов, а также полученные результаты могут использоваться физическими лицами при принятии решений о купле/продаже недвижимости.

BaseGroup Labs – профессиональный поставщик продуктов и решений в области анализа данных. Мы имеем многолетний опыт работы в области разработки аналитических алгоритмов и создания законченных систем. BaseGroup Labs предлагает полностью интегрированные продукты, объединяющие все необходимые инструменты анализа: хранилища данных, аналитическую отчетность, механизмы поиска закономерностей и построения моделей, средства интеграции аналитических систем с платформами сторонних производителей.

Системы от BaseGroup Labs выполнены с применением **самых современных** информационных технологий.

Технологии

Data Warehouse
хранилище данных

- Консолидация анализируемых данных, обеспечение непротиворечивости данных
- Быстрый доступ к необходимой информации
- Автоматическое обновление данных
- Богатый семантический слой

Data Mining
добыча данных

- Прогнозирование
- Поиск закономерностей и зависимостей
- Извлечение правил
- Оптимизация процессов
- Анализ по принципу «что-если»

OLAP
многомерный анализ данных

- Многомерная отчетность, позволяющая извлечь максимум полезной информации из имеющихся данных
- Гибкие механизмы навигации и манипулирования данными
- Анализ тенденций
- Простота использования конечным пользователем

Knowledge Discovery in DB
обнаружение знаний в базах данных

- Механизмы улучшения качества исходных данных (очистка, преобразование и трансформация данных)
- Построение сценариев обработки данных
- Механизмы построения моделей
- Интеграция моделей в информационные системы

Наши системы базируются на **собственном аналитическом ядре**, что обеспечивает беспрецедентную **гибкость** при выборе способов анализа и создании прикладных решений. Применение самообучающихся механизмов дает возможность быстрой адаптации решения под постоянно изменяющиеся условия.

Россия, 390046, г. Рязань, Введенская, д. 115, оф. 447
Т./ф.: +7 (4912) 24-09-77, +7 (4912) 24-06-99
info@basegroup.ru
www.basegroup.ru

Deductor – флагманский продукт BaseGroup Labs, концентрирующий многолетний опыт компании и вобравший в себя самые удачные архитектурные идеи и современный математический аппарат. В Deductor реализованы технологии анализа структурированных данных: нейронные сети, деревья решений, хранилища данных и OLAP, ассоциативные правила, карты Кохонена и многое другие. Использование Deductor в учебном процессе поможет студентам освоить алгоритмы машинного обучения и системы интеллектуальной обработки информации на практике, решая актуальные задачи по консолидации, очистке, прогнозированию, классификации, кластеризации, скорингу.

Образование

Для высших учебных заведений BaseGroup Labs предлагает специальные условия. Заключив с нами соглашение о сотрудничестве, преподаватели и сотрудники учебного заведения получают следующие возможности:

- Аналитическую платформу **Deductor Academic** для проведения практикумов по дисциплинам, связанным с информационно-аналитическими системами, интеллектуальными информационными системами, системами поддержки принятия решений и другим курсам для прикладных информатиков и экономистов.
- Бесплатное e-learning обучение преподавателей на образовательном портале **edu.basegroup.ru** в полноценной системе **дистанционного обучения** и сертификацию по результатам обучения, обсуждение возникающих вопросов на форуме.
- Программы **повышения квалификации** по корпоративным аналитическим системам при Институте Экономики и Финансов «Синергия».
- Большое число **методических разработок** для проведения практических занятий со студентами по всем современным технологиям анализа данных.

Участие в программе полностью **бесплатное**. Образовательная инициатива действует с 2005 года и за это время более **70** вузов России, Украины и Беларуси стали нашими партнерами и используют аналитическую платформу Deductor в учебном процессе. Вот некоторые из них:

- Российская экономическая академия имени Г.В. Плеханова;
- Государственный университет управления;
- Московский авиационный институт;
- Санкт-Петербургский государственный университет;
- Белорусский государственный университет информатики и радиоэлектроники.

Форму и условия партнерства, полный список вузов-партнеров и другую дополнительную информацию можно получить на образовательном портале **http://edu.basegroup.ru**.

Россия, 390046, г. Рязань, Введенская, д. 115, оф. 447
Т./ф.: +7 (4912) 24-09-77
+7 (4912) 24-06-99
education@basegroup.ru

БИЗНЕС-АНАЛИТИКА. ВОПРОСЫ ТЕОРИИ И ПРАКТИКИ. ИСПОЛЬЗОВАНИЕ АНАЛИТИЧЕСКОЙ ПЛАТФОРМЫ DEDUCTOR В ДЕЯТЕЛЬНОСТИ УЧЕБНЫХ ЗАВЕДЕНИЙ:

Сборник материалов межвузовской научно-практической конференции

24 июня 2010 года

Тираж 100 экз. Подписано в печать 27.09.2010 г.
Формат 60x84 1/16. Гарнитура «Times New Roman».
Усл. печ. л. Уч.-изд. л.

Отпечатано с готовых диапозитивов
в ООО фирма «Интермета»
г. Рязань, ул. Семинарская, 5, т. 25-81-76