МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
Запорізький національний технічний університет

Інформаційні системи та технології в управлінні

МЕТОДИЧНІ ВКАЗІВКИ

теоретичні відомості і завдання до лабораторних робіт

для студентів та магістрів денної форми навчання
спеціальності 7.803060101
Менеджмент організацій і адміністрування
Частина 1
Прогнозування часових рядів

2014

Інформаційні системи та технології в управлінні. Методичні вказівки, теоретичні відомості і завдання до лабораторних робіт для студентів та магістрів денної форми навчання спеціальності 7.803060101 Менеджмент організацій і адміністрування. Частина 1. Прогнозування часових рядів. / Укл.: Біла Н.І. – Запоріжжя: ЗНТУ, 2014. – с. 60.

Містить теоретичні відомості, індивідуальні завдання та приклади із курсу «Інформаційні системи та технології в управлінні» за темою «Прогнозування часових рядів».

Укладачі: Біла Н.І. доцент,

Рецензенти: Пінчук В.П., доцент
Вишневська В.Г., доцент.

Відповідальний за випуск Корніч Г.В., зав. кафедрою, професор

Затверджено на засіданні кафедри
обчислювальної математики,
протокол № 2 від 28.03.2014

ЗМІСТ

1 ВВЕДЕНИЕ В СИСТЕМЫ ПОДДЕРЖКИ ПРИНЯТИЯ РЕШЕНИЙ (СППР)

1.1 Определение СППР

Информационные системы являются в наше время неотъемлемой частью технологий управления бизнесом. Практически на каждом предприятии работают информационные системы, осуществляющие функции учета и контроля деятельности фирмы. Со структурой и принципами работы таких систем вы знакомились в курсе «Информационные системы менеджмента». Однако, существуют информационные системы другого типа, которые называют системы поддержки принятия решений (СППР).

СППР возникли в результате развития управленческих информационных систем и систем управления базами данных в начале 70-х годов прошлого века. На данный момент существует огромное количество СППР, разработанных и внедренных в различных областях человеческой деятельности. Темпы их разработок постоянно возрастают.

СППР - интерактивная компьютерная система, предназначенная для поддержки принятия решений в слабоструктурированных и неструктурированных проблемах различных видов человеческой деятельности.

Существенными концепциями этого определения являются:

- компьютерная интерактивная;
- поддержка принятия решений (решение принимает человек);
- слабоструктурированных и неструктурированных проблем (именно такими проблемами занимаются руководители).

Рассмотрим, что же представляет собой классификация проблем на слабо структурированные, неструктурированные и структурированные.

Неструктурированные задачи имеют только качественное описание, основанное на суждениях ЛПР (лица принимающего решения), количественные зависимости между основными характеристиками задачи не известны.

Структурированные задачи характеризуются существенными зависимостями, которые могут быть выражены количественно.

Слабоструктурированные задачи занимают промежуточное положение и являются "сочетающими количественные и качественные зависимости, причем малоизвестные и неопределенные стороны задачи имеют тенденцию доминировать".

Можно выделить три компонента, составляющие основу классической структуры СППР, которыми она отличается от других типов информационных систем: подсистему интерфейса пользователя, подсистему управления базой данных и подсистему управления базой моделей.

Если посмотреть на СППР с функциональной стороны, можно выделить следующие ее компоненты:

- сервер хранилища данных;
- инструментарий OLAP;
- инструментарий Data Mining.

Эти компоненты СППР рассматривают такие основные вопросы: вопрос накопления данных и их моделирования на концептуальном уровне, вопрос эффективной загрузки данных из нескольких независимых источников и вопрос анализа данных.

Можно сказать, что использование оперативной аналитической обработки (систем OLAP) на сегодня ограничивается обеспечением доступа к многомерным данным.

Технология Data Mining представляет в СППР наибольший интерес, поскольку с ее помощью можно провести наиболее глубокий и всесторонний анализ данных и, следовательно, принимать наиболее взвешенные и обоснованные решения.

В той или иной степени Системы Поддержки Принятия Решений (СППР) присутствуют в любой информационной системе. Поэтому, осознанно или нет, к задаче создания системы поддержки принятия решений организации приступают сразу после приобретения вычислительной техники и установки программного обеспечения. По мере развития бизнеса, упорядочения структуры организации и

налаживания межкорпоративных связей, проблема разработки и внедрения СППР становится особенно актуальной.

1.2 Классификация СППР

СППР можно, в зависимости от данных, с которыми они работают, разделить на:

- оперативные - предназначены для немедленного реагирования на текущую ситуацию;
- стратегические - основанные на анализе большого количества информации из разных источников с привлечением сведений, содержащихся в системах, аккумулирующих опыт решения проблем.

СППР первого типа получили название Информационных Систем Руководства (Executive Information Systems, ИСР). По сути, они представляют собой конечные наборы отчетов, построенные на основании данных из транзакционной информационной системы предприятия или OLTP-системы, в идеале адекватно отражающей в режиме реального времени все аспекты производственного цикла предприятия.

СППР второго типа - Decision Support System (DSS) предполагают достаточно глубокую проработку данных, специально преобразованных так, чтобы их было удобно использовать в ходе процесса принятия решений. Неотъемлемым компонентом СППР этого уровня являются правила принятия решений, которые на основе агрегированных данных подсказывают менеджерскому составу выводы и придают системе черты искусственного интеллекта. Такого рода системы создаются только в том случае, если структура бизнеса уже достаточно определена и имеются основания для обобщения и анализа не только данных, но и процессов их обработки. Если ИСР есть не что иное как развитие системы оперативного управления производственными процессами, то СППР в современном понимании - это механизм развития бизнеса, который включает в себя некоторую часть управляющей информационной системы, обширную систему внешних связей предприятия, а также технологические и маркетинговые процессы развития производства.

СППР имеет смысл создавать, если есть основания для обобщения и анализа данных и процессов их обработки. Системы этого типа иногда называют динамическими, т.е. они должны быть ориентированы на обработку неожиданных (ad hoc) запросов.

1.3 Архитектура СППР

Рисунок 1 - Обобщенная архитектура системы поддержки принятия решений

Поддержка принятия решений на основе накопленных данных может выполняться в трех базовых сферах.

1. Область детализированных данных (OLTP-системы). Целью большинства таких систем является поиск информации, это так называемые информационно-поисковые системы. Они могут использоваться в качестве надстроек над системами обработки данных или как хранилища данных.

2. Сфера агрегированных показателей (OLAP-системы). Задачами OLAP систем является обобщение, агрегация, гиперкубическое представление информации и многомерный анализ. Это могут быть многомерные СУБД или же реляционные базы с предварительной агрегацией данных.

3. Сфера закономерностей (Data Mining).

Общая схема поддержки принятия решений включает:

- помощь ЛПР при оценке состояния управляемой системы и воздействий на нее; выявление предпочтений ЛПР;
- генерацию возможных решений;
- оценку возможных альтернатив, исходя из предпочтений ЛПР;
- анализ последствий принимаемых решений и выбор лучшего с точки зрения ЛПР.

Системы рассчитаны на пользователей, имеющих как знания в предметной области, так и возможности использования современных компьютерных технологий. Этим системам присущи черты искусственного интеллекта, за счет возможности проработки исходных данных в конкретные выводы по поставленной задаче.

1.4 Анализ данных – основные принципы

Анализ информации является неотъемлемой частью ведения бизнеса и одним из важных факторов повышения его конкурентоспособности. При этом в подавляющем большинстве случаев анализ сводится к применению одних и тех же базовых механизмов. Они являются универсальными и применимы к любой предметной области, благодаря чему имеется возможность создания унифицированной программной платформы, в которой реализованы основные механизмы анализа.

Обычно анализ производят аналитики и эксперты предметной области предприятия. Они подготавливают данные к пригодному для анализа виду, применяют к ним различные методы анализа, приводят результаты к легко воспринимаемому виду. Результаты анализа необходимы лицам предприятия, принимающим решения, например, руководителям отделов, менеджерам.

Таким образом, требуется, с одной стороны, выделить и формализовать знание эксперта о предметной области, с другой, обеспечить возможность использовать эти знания человеком, не разбирающимся в особенностях использования механизмов анализа, т.е. решить проблему тиражирования знаний.

Любая организация в процессе своей деятельности стремится повысить прибыль и уменьшить расходы. В этом ей помогают новые компьютерные технологии, использование разнообразных программ

автоматизации бизнес-процессов. Это учетные, бухгалтерские и складские системы, системы управленческого учета и многие другие. Чем аккуратнее и полнее ведется сбор и систематизация информации, тем полнее будет представление о процессах в организации.

Современные носители информации позволяют хранить десятки и сотни гигабайт информации, но без использования специальных средств анализа накопленной информации такие носители превращаются просто в свалку бесполезных сведений. Очень часто принятие правильного решения затруднено тем, что хотя данные и имеются, они являются неполными, или, наоборот, избыточными, замусорены информацией, которая вообще не имеет отношения к делу, несистематизированными или систематизированными неверно. Тогда прибегают к помощи программных средств, которые позволяют привести информацию к виду, который дает возможность с достаточной степенью достоверности оценить содержащиеся в ней факты и повысить вероятность принятия оптимального решения.

Есть ***два подхода*** к анализу данных с помощью информационных систем.

В первом варианте программа используется для визуализации информации - извлечения данных из источников и предоставления их человеку для самостоятельного анализа и принятия решений. Обычно данные, предоставляемые программой, являются простой таблицей, и в таком виде их очень сложно анализировать, особенно если данных много, но имеются и более удобные способы отображения: кросс-таблицы, диаграммы, гистограммы, карты, деревья. Этот вариант анализа данных реализован в OLAP технологии.

Второй вариант использования программного обеспечения для анализа – это построение моделей. Модель имитирует некоторый процесс, например, изменение объемов продаж некоторого товара, поведение клиентов и другое. Для построения модели необходимо сделать предобработку данных и далее к ним применять математические методы анализа: кластеризацию, классификацию, регрессию и т. д. Построенную модель можно использовать для принятия решений, объяснения причин, оценки значимости факторов, моделирования различных вариантов развития. Этот вариант анализа данных реализуется в системах Data Mining.

Как визуализация, так и построение моделей осуществляются путем применения к данным базовых методов анализа. Это достаточно известные методы, и они используются в самых разнообразных сферах деятельности.

1.5 Базовые методы анализа

1) Online Analytical Processing

Любая система поддержки принятия решений, прежде всего, должна обладать средствами отбора и предоставления пользователю данных в удобной для восприятия и анализа форме. Как правило, наиболее удобными для анализа являются многомерные данные, описывающие предметную область сразу с нескольких точек зрения. Для описания таких наборов данных вводится понятие многомерных кубов (гиперкубов, метакубов). По осям такого куба размещаются параметры - измерения, а в ячейках – зависящие от них данные - факты. Вдоль каждой оси представлены различные уровни детализации данных. Использование такой модели данных позволяет повысить эффективность работы с ними: генерировать сложные запросы, создавать отчеты, выделять подмножества данных и т.д. Технология комплексного многомерного анализа данных и предоставления результатов этого анализа в удобной для использования форме получила название OLAP.

OLAP (Online Analytical Processing) – оперативная аналитическая обработка данных. OLAP дает возможность в реальном времени генерировать описательные и сравнительные сводки данных и получать ответы на различные другие аналитические запросы. OLAP-кубы представляют собой проекцию исходного куба данных на куб данных меньшей размерности. При этом значения ячеек агрегируются, то есть объединяются с применением функции агрегации – сумма, среднее, количество, минимум, максимум. Такие проекции или срезы исходного куба представляются на экране в виде кросс-таблицы.

2) Knowledge Discovery in Databases

KDD (Knowledge Discovery in Databases) – извлечение знаний из баз данных. Это процесс поиска полезных знаний в «сырых данных». KDD включает в себя вопросы подготовки данных, выбора информативных признаков, очистки данных, применения методов

Data Mining (DM), постобработки данных и интерпретации полученных результатов.

Привлекательность этого подхода заключается в том, что вне зависимости от предметной области мы применяем одни и те же операции:

1. ***Подготовка исходного набора данных.*** Этот этап заключается в создании набора данных, в том числе слиянии сведений из различных источников, определение выборки, которая и будет впоследствии анализироваться. Для этого должны существовать развитые инструменты доступа к различным источникам данных: файлам разных форматов, базам данных, учетным системам.

2. ***Предобработка и очистка данных***. Для того чтобы эффективно применять методы анализа, следует обратить серьезное внимание на вопросы предобработки данных. Данные могут содержать пропуски, шумы, аномальные значения и т.д. Кроме того, данные могут быть избыточны, недостаточны и т.д. В некоторых задачах требуется дополнить данные некоторой априорной информацией. Наивно предполагать, что если подать любые данные на вход системы в существующем виде, то на выходе получим полезные знания. Данные должны быть качественны и корректны с точки зрения используемого метода анализа. Более того, иногда размерность исходного пространства может быть очень большой, и тогда желательно применение специальных алгоритмов понижения размерности: наиболее значимых признаков и отображение данных в пространство меньшей размерности.

3. ***Трансформация данных***. Для различных методов анализа требуются данные, подготовленные в специальном виде. Например, некоторые методы анализа в качестве входных полей могут использовать только числовые данные, а некоторые, наоборот, только категориальные.

4. ***Data Mining***. На этом шаге применяются различные алгоритмы для нахождения знаний. Это нейронные сети, деревья решений, алгоритмы кластеризации и установления ассоциаций и т.д. Для этого могут использоваться как классические статистические методы, так и самообучающиеся алгоритмы и машинное обучение.

5. ***Постобработка данных***. Тестирование, интерпретация результатов и практическое применение полученных знаний в бизнесе.

Описанный процесс повторяется итеративно, а реализация этих этапов позволяет автоматизировать процесс извлечения знаний.

Например, нужно сделать прогноз объемов продаж на следующий месяц. Есть сеть магазинов розничной торговли. Первым шагом будет сбор истории продаж в каждом магазине и объединение ее в общую выборку данных. Следующим шагом будет предобработка собранных данных. Например, их группировка по месяцам, сглаживание кривой продаж, устранение факторов, слабо влияющих на объемы продаж. Далее следует построить модель зависимости объемов продаж от выбранных факторов. Это можно сделать с помощью линейной регрессии или нейронных сетей.

Имея такую модель, можно получить прогноз, подав на вход модели нашу историю продаж. Зная прогнозное значение, его можно использовать, например, для оптимизации закупок товара.

3) Data Mining

DM (Data Mining) – «добыча» данных. Это метод обнаружения в «сырых» данных ранее неизвестных, нетривиальных, практически полезных и доступных для интерпретации знаний, необходимых для принятия решений в различных сферах человеческой деятельности. DM обеспечивает решение всего пяти задач — классификация, кластеризация, регрессия, ассоциация, последовательность:

1. ***Классификация*** — установление функциональной зависимости между входными и дискретными выходными переменными. При помощи классификации решается задача отнесение объектов (наблюдений, событий) к одному из заранее известных классов.

2. ***Регрессия*** – установление функциональной зависимости между входными и непрерывными выходными переменными. Прогнозирование чаще всего сводится к решению задачи регрессии.

3. ***Кластеризация*** — это группировка объектов (наблюдений, событий) на основе данных (свойств), описывающих сущность объектов. Объекты внутри кластера должны быть «похожими» друг на

друга и отличаться от объектов, вошедших в другие кластеры. Чем больше похожи объекты внутри кластера и чем больше отличий между кластерами, тем точнее кластеризация.

4. ***Ассоциация*** — выявление зависимостей между связанными событиями, указывающих, что из события X следует событие Y. Такие правила называются ассоциативными. Впервые эта задача была предложена для нахождения типичных шаблонов покупок, совершаемых в супермаркетах, поэтому иногда ее еще называют анализом потребительской корзины (market basket analysis).

5. ***Последовательные шаблоны*** — установление закономерностей между связанными во времени событиями. Например, после события X через определенное время произойдет событие Y.

6. Иногда специально выделяют ***задачу анализа отклонений*** — выявление наиболее нехарактерных шаблонов.

1.6 Примеры задач, где применяются методы Data Mining

Классификация используется в случае, если заранее известны классы отнесения объектов. Например, отнесение нового товара к той или иной товарной группе, отнесение клиента к какой-либо категории. При кредитовании это может быть, например, отнесение клиента по каким-то признакам к одной из групп риска.

Регрессия чаще всего используется при прогнозировании объемов продаж. В этом случае зависимой величиной являются объемы продаж, а факторами, влияющими на эту величину, могут быть предыдущие объемы продаж, изменение курса валют, активность конкурентов и т.д. или, например, при диагностике оборудования, когда оценивается зависимость надежности от различных внешних факторов, показателей датчиков, износа оборудования.

Кластеризация может использоваться для сегментирования и построения профилей клиентов (покупателей). При достаточно большом количестве клиентов становится трудно подходить к каждому индивидуально. Поэтому клиентов удобно объединить в группы – сегменты со сходными признаками. Выделять сегменты клиентов можно по нескольким группам признаков. Это могут быть сегменты по сфере деятельности, по географическому расположению.

После сегментации можно узнать, какие именно сегменты являются наиболее активными, какие приносят наибольшую прибыль, выделить характерные для них признаки. Эффективность работы с клиентами повышается за счет учета их персональных или групповых предпочтений.

Ассоциации помогают выявлять совместно приобретаемые товары. Это может быть полезно для более удобного размещения товара на прилавках, стимулирования продаж. Тогда человек, купивший пачку спагетти, не забудет купить к ним бутылочку соуса.

Последовательные шаблоны могут быть использованы, например, при планировании продаж или предоставлении услуг. Например, если человек приобрел фотопленку, то через неделю он отдаст ее на проявку и закажет печать фотографий.

Для анализа отклонений необходимо сначала построить шаблон типичного поведения изучаемого объекта. Например, поведение человека при использовании кредитных карт. Тогда будет известно, что клиент (покупатель) использует карту регулярно два раза в месяц и приобретает товар в пределах определенной суммы. Отклонением будет, например, не запланированное приобретение товара по данной карте на большую сумму. Это может говорить об ее использовании другим лицом, то есть о факте мошенничества.

Следует отметить, что на сегодняшний день наибольшее распространение технология Data Mining получила при решении бизнес-задач. Возможно, причина в том, что именно в этом направлении отдача от использования инструментов Data Mining может составлять, по некоторым источникам, до 1000% и затраты на ее внедрение могут достаточно быстро окупиться.

Сейчас технология Data Mining используется практически во всех сферах деятельности человека, где накоплены ретроспективные данные. Назовем часть из них:

1. Применение Data Mining для решения бизнес-задач. Основные направления: банковское дело, финансы, страхование, CRM, производство, телекоммуникации, электронная коммерция, маркетинг, фондовый рынок и другие.

2. Применение Data Mining для решения задач государственного уровня. Основные направления: поиск лиц, уклоняющихся от налогов; средства в борьбе с терроризмом.
3. Применение Data Mining для научных исследований. Основные направления: медицина, биология, молекулярная генетика и генная инженерия, биоинформатика, астрономия, прикладная химия, исследования, касающиеся наркотической зависимости, и другие.
4. Применение Data Mining для решения Web-задач. Основные направления: поисковые машины (search engines), счетчики и другие.

Одно из наиболее перспективных направлений применения Data Mining - использование данной технологии в аналитическом CRM.

CRM (Customer Relationship Management) - управление отношениями с клиентами. При совместном использовании этих технологий добыча знаний совмещается с "добычей денег" из данных о клиентах.

Важным аспектом в работе отделов маркетинга и отдела продаж является составление целостного представления о клиентах, информация об их особенностях, характеристиках, структуре клиентской базы. В CRM используется так называемое профилирование клиентов, дающее полное представление всей необходимой информации о клиентах. Профилирование клиентов включает следующие компоненты: сегментация клиентов, прибыльность клиентов, удержание клиентов, анализ реакции клиентов. Каждый из этих компонентов может исследоваться при помощи Data Mining, а анализ их в совокупности, как компонентов профилирования, в результате может дать те знания, которые из каждой отдельной характеристики получить невозможно.

В результате использования Data Mining решается задача сегментации клиентов на основе их прибыльности. Анализ выделяет те сегменты покупателей, которые приносят наибольшую прибыль. Сегментация также может осуществляться на основе лояльности клиентов. В результате сегментации вся клиентская база будет поделена на определенные сегменты, с общими характеристиками. В

соответствии с этими характеристиками компания может индивидуально подбирать маркетинговую политику для каждой группы клиентов.

Также можно использовать технологию Data Mining для прогнозирования реакции определенного сегмента клиентов на определенный вид рекламы или рекламных акций - на основе ретроспективных данных, накопленных в предыдущие периоды.

Таким образом, определяя закономерности поведения клиентов при помощи технологии Data Mining, можно существенно повысить эффективность работы отделов маркетинга, продаж и сбыта. При объединении технологий CRM и Data Mining и грамотном их внедрении в бизнес компания получает значительные преимущества перед конкурентами.

Перечисленные выше базовые методы анализа данных используются для создания аналитических систем. Причем, под такой системой понимается не только какая-то одна программа. Некоторые механизмы анализа могут быть реализованы на бумаге, некоторые на компьютере с использованием электронных таблиц, баз данных и других приложений. Однако, такой подход при частом использовании не эффективен. Намного лучшие результаты даст применение единого хранилища данных и единой программы, содержащей в себе всю функциональность, необходимую для реализации концепции KDD.

1.7 Программа Deductor – платформа для создания СППР

Deductor предназначен для эффективного решения проблемы тиражирования знаний, это аналитическая платформа, основа для создания законченных прикладных решений в области анализа данных. Deductor состоит из двух компонентов: аналитического приложения Deductor Studio и многомерного хранилища данных Deductor Warehouse.

Deductor Warehouse - многомерное хранилище данных, аккумулирующее всю необходимую для анализа предметной области информацию. Использование единого хранилища позволяет обеспечить непротиворечивость данных, их централизованное хранение и автоматически создает всю необходимую поддержку процесса анализа данных. Deductor Warehouse оптимизирован для

решения именно аналитических задач, что положительно сказывается на скорости доступа к данным.

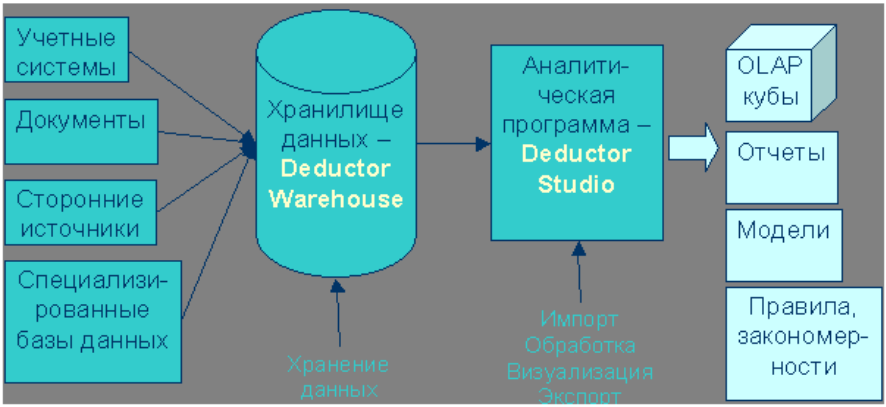


Рисунок 1.2 - Архитектура системы Deductor

Deductor Studio - это программа, предназначенная для анализа информации из различных источников данных. Она реализует функции импорта, обработки, визуализации и экспорта данных. Deductor Studio может функционировать и без хранилища данных, получая информацию из любых других источников, но наиболее оптимальным является их совместное использование.

Поддержка процесса от разведочного анализа до отображения данных Deductor Studio позволяет пройти все этапы анализа данных.

В лабораторных работах вы ознакомитесь с методами анализа бизнес – информации, некоторыми практическими задачами анализа и способами их решения с использованием программы Deductor Academic. Программа распространяется бесплатно и, в первую очередь, предназначена для обучения.

Последовательность действий, которые необходимо провести для анализа данных, называется сценарием. Сценарий можно автоматически выполнять на любых данных.

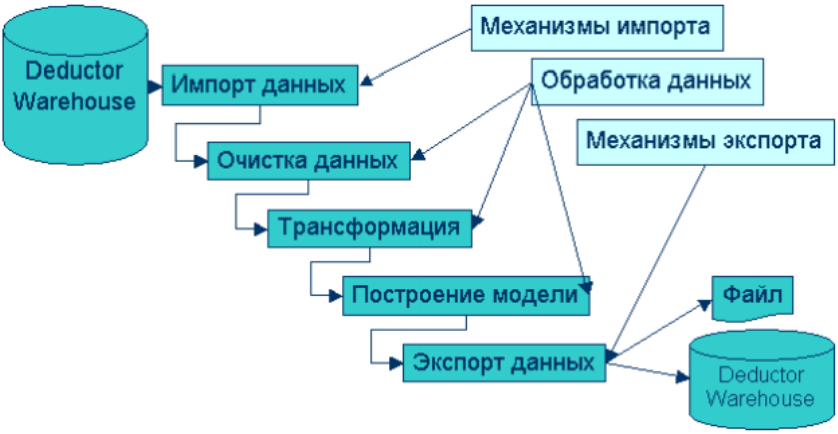


Рисунок 1.3 – Типовой сценарий анализа данных в Deductor

1.8 Контрольные вопросы

1. Дайте определение СППР.
2. Назовите главные составляющие СППР.
3. Приведите примеры использования интеллектуального анализа данных в бизнесе и, в частности, в менеджменте.
4. Какие процессы обозначают термином KDD?
5. Какие задачи решаются методами Data Mining?
6. Приведите примеры бизнес-приложений, в которых используются методы Data Mining.
7. К какому классу программ относится программа Deductor?
8. Какое еще программное обеспечение используется для решения задач Data Mining?
9. Какие два подхода к анализу данных Вы можете назвать?

2 КОРРЕЛЯЦИОННЫЙ АНАЛИЗ

2.1 Теоретические сведения

Корреляционный анализ применяется для оценки зависимости выходных полей данных от входных факторов и устранения незначащих факторов. Принцип корреляционного анализа состоит в поиске таких значений, которые в наименьшей степени коррелированны (взаимосвязаны) с выходным результатом. Такие факторы могут быть исключены из результирующего набора данных практически без потери полезной информации. Критерием принятия решения об исключении является порог значимости. Если модуль корреляции (степень взаимозависимости) между входным и выходным факторами меньше порога значимости, то соответствующий фактор отбрасывается как незначащий.

В процессе обработки значащие факторы могут выбираться вручную или автоматически. При ручном выборе около имени каждого входного поля устанавливается флажок, если это поле нужно включить в выходную выборку, и снимается в противном случае. В автоматическом режиме исключаются все факторы, корреляция которых с выходными полями меньше порога задаваемого уровня значимости.

Корреляционный анализ применяется для количественной оценки взаимосвязи двух наборов данных, представленных в безразмерном виде. Коэффициент корреляции, всегда обозначаемый латинской буквой r, используется для определения наличия взаимосвязи между двумя свойствами.

Связь между признаками (по шкале Чеддока) может быть сильной, средней и слабой. Тесноту связи определяют по величине коэффициента корреляции, который может принимать значения от -1 до +1 включительно.

Таблица 2.1 – Критерии оценки тесноты связи

Величина коэффициента корреляции	0,1 – 0,3	0,3 – 0,5	0,5 – 0,7	0,7 – 0,9	0,9 – 1,0
Характеристика силы связи	слабая	умеренная	заметная	высокая	весьма высокая

Пример 2.1.

В качестве примера рассмотрим, как определить товары-заменители и сопутствующие товары, имея временные ряды объемов продаж. У товаров-заменителей должна быть большая отрицательная корреляция, т.к. увеличение продаж одного товара ведет к спаду продаж второго. А у сопутствующих товаров – большая положительная корреляция.

Пусть есть такие временные ряды продаж товаров:

Таблица 2.1

Товар1	**Товар2**	**Товар3**	**Товар4**
10	20	15	25
12	22	12	26
14	25	9	26
13	24	10	25
14	25	9	24
14	25	9	23
12	21	12	24
10	18	14	23
16	24	9	22
13	21	9	23
17	25	7	25

Определим корреляцию Товар1 с остальными товарами. Данные о продажах находятся в файле товар.txt.

Для решения задачи будем использовать программу Deductor.

На первом шаге решения задачи нужно загрузить в Deductor данные из текстового файла. Для этого в левом окне программы Deductor нажимаем кнопку «Мастер импорта».

Импорт данных осуществляется в режиме диалога, вам нужно только правильно отвечать на вопросы мастера.

На первом шаге укажите, что данные будут читаться из текстового файла (Text), и и укажите имя файла. Файл можно выбрать, используя кнопку с многоточием (…). Результат представлен на рисунке 2.1.

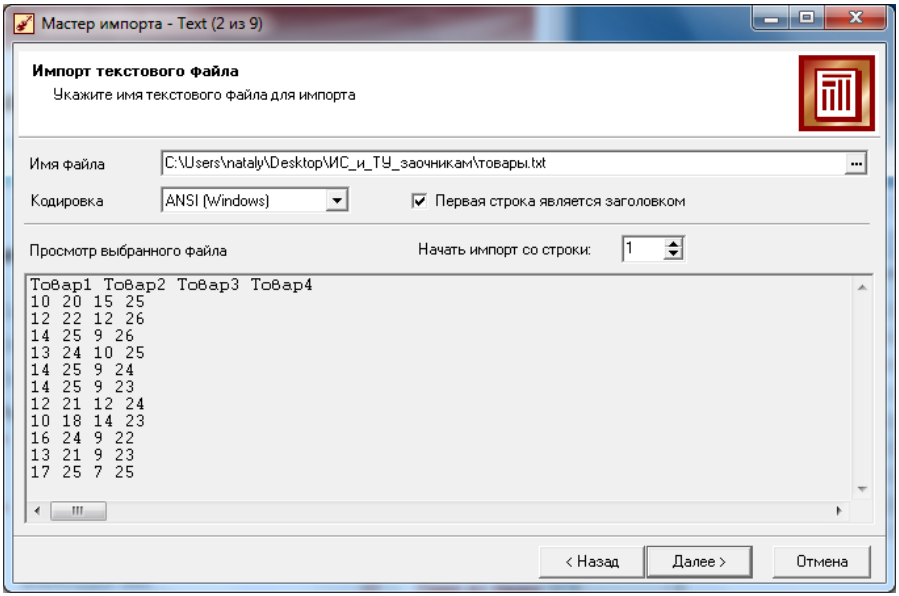


Рисунок 2.1 – Ввод данных из файла

На третьем шаге мастера импорта выбираем переключатель «С разделителями». Поскольку данные в текстовом файле отделены друг от друга пробелами, на следующем шаге указываем, что разделителем является пробел.

На следующем шаге указываем типы данных в столбцах. Deductor определяет тип данных автоматически, вам нужно проверить, правильно ли определены типы данных и откорректировать их в случае необходимости. Результат работы на этом шаге представлен на рис. 2.2.

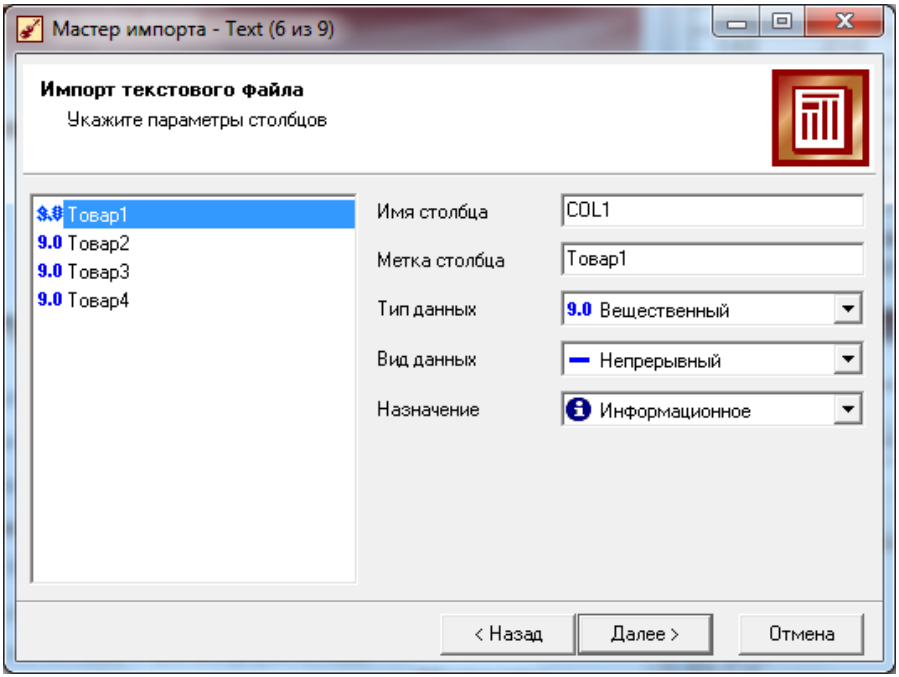


Рисунок 2.2 – Определение параметров столбцов

На следующем шаге нажмите кнопку «Пуск», чтобы запустить процесс загрузки файла. Затем укажите способ отображения данных как показано на рис. 2.3.

На рис. 2.4 показан результат загрузки данных и отображение их в виде таблицы.

Теперь можно приступить к обработке данных. Для этого вызываем «Мастер обработки» и выбираем пункт «Корреляционный анализ», как показано на рис. 2.5.

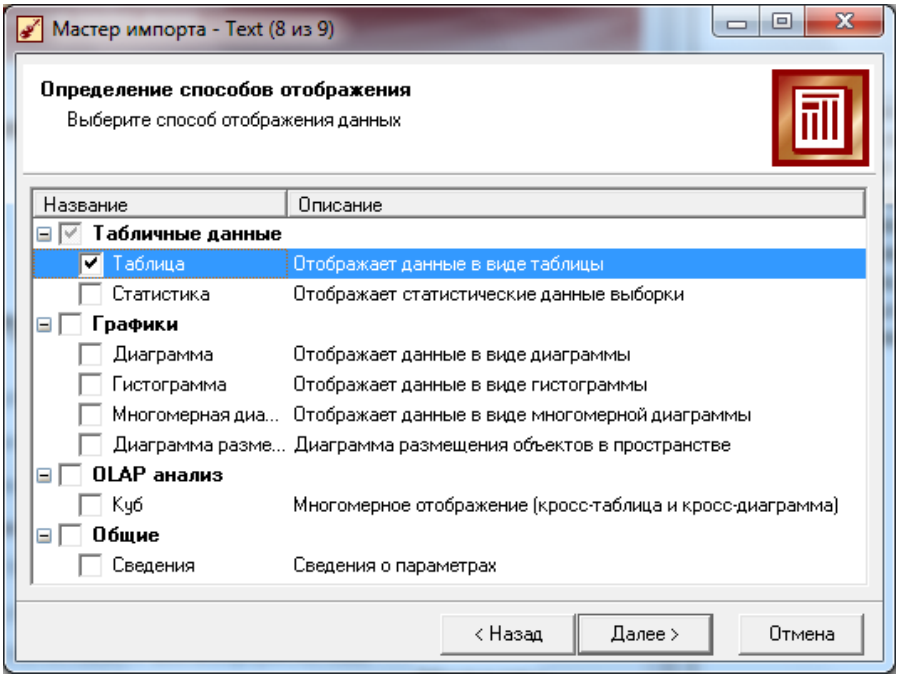
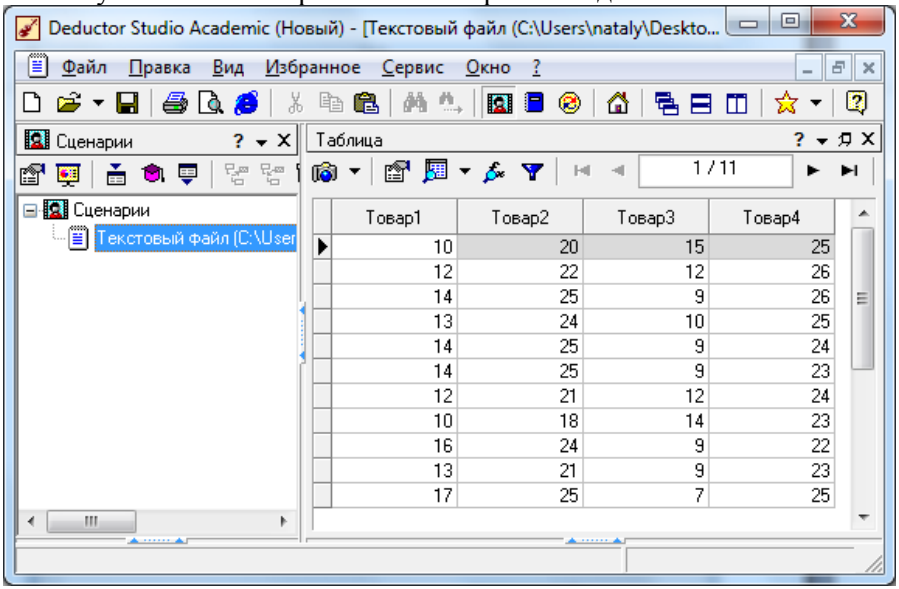


Рисунок 2.3 – Выбор способа отображения данных



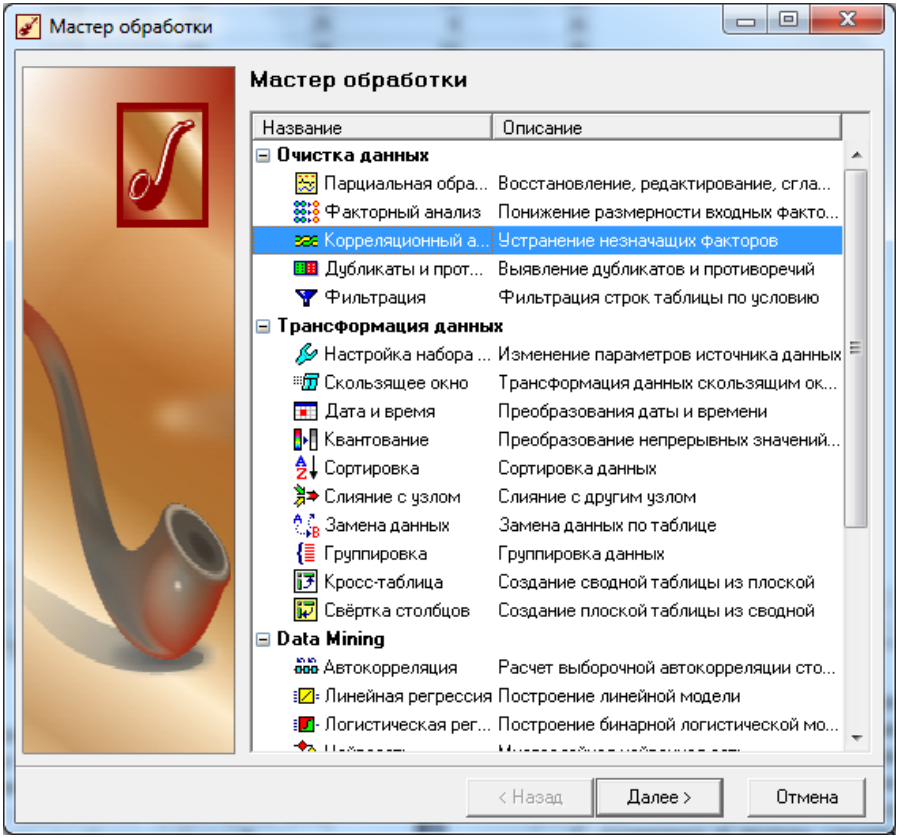


Рисунок 2.5 – Выбор метода обработки данных

На первом шаге корреляционного анализа нужно определить какие данные являются входными, а какие выходными. Также можно указать, какие данные не будут использоваться при анализе. В этом случае они могут быть информационными или неиспользуемыми.

Поскольку мы хотим определить степень зависимости между продажами Товара1 и остальных товаров, то указываем Товар1 как выходной, а остальные товары входными, как показано на рис. 2.6.

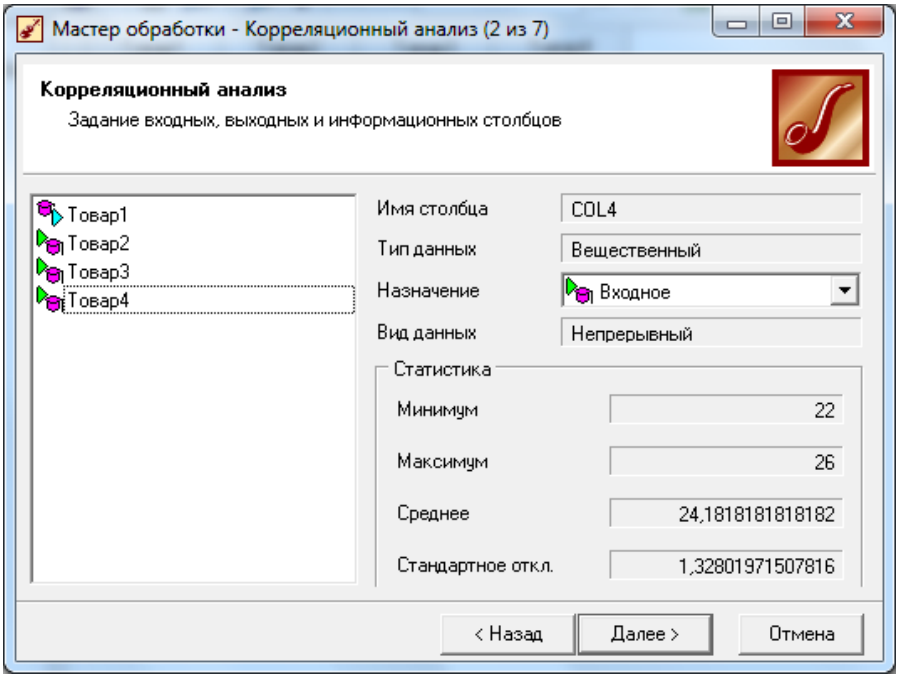


Рисунок 2.6 – Задание входных и выходных столбцов для корреляционного анализа.

На следующем шаге выбираем «Коэффициент корреляции Пирсона», а затем нажимаем кнопку «Пуск», чтобы запустить процесс вычисления коэффициентов корреляции.

На следующем шаге, когда коэффициенты корреляции посчитаны, можно отбирать значащие факторы. Это можно сделать вручную или автоматически. В последнем случае необходимо указать порог значимости. На рисунке 2.7 указан очень низкий порог значимости, поэтому отбираются все переменные.

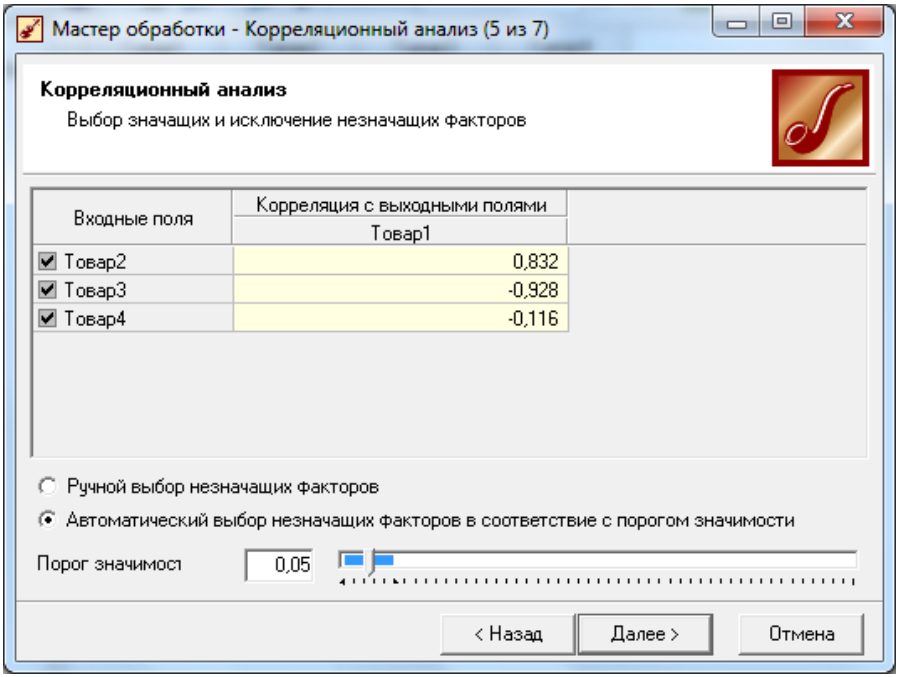


Рисунок 2.7 – Выбор значащих факторов

Одним из доступных способов визуализации результатов является визуализатор «Матрица корреляции». В данном примере эта матрица имеет следующий вид:

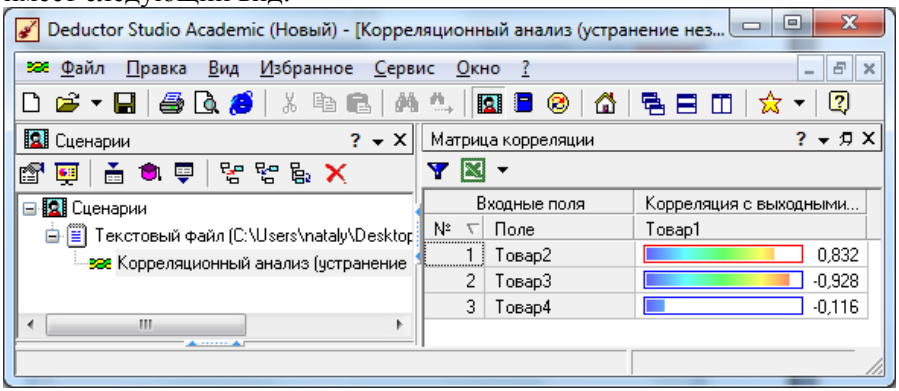


Рисунок 2.8 – Результат корреляционного анализа

Как видно из рисунка 2.7, ряд продаж для Товар2 имеет очень большую положительную, а Товар3 – отрицательную корреляцию. Из этого можно сделать вывод, что Товар2, возможно, является сопутствующим товаром, а Товар3 – заместителем Товар1. Корреляция с продажами Товар4 Товара1 является отрицательной, но при этом абсолютное значение корреляции невелико, и, следовательно, можно говорить об отсутствии взаимосвязи между продажами Товар1 и продажами Товар4.

2.3 Задание для самостоятельной работы

Загрузите данные из файла "region.txt". В данном примере необходимо определить степень влияния экономических показателей региона на среднедушевой денежный доход жителей. Укажите, какие показатели влияют на денежный доход, а какие можно отбросить.

2.4. Контрольные вопросы

1. Какие задачи решают с помощью корреляционного анализа?
2. Постановка задачи отбора значимых факторов (что дано, что хотим получить).
3. Как определяют незначимые факторы?
4. Что такое коэффициент корреляции?
5. Какие значения может принимать коэффициент корреляции?
6. О чем говорит значеник коэффициента корреляции 0,86?
7. Как оценивают связь между признаками в таблице с данными?
8. Каким является первый шаг при использовании программы Deductor?
9. Какой обработчик используют в программе Deductor, чтобы определить значимые факторы?

3 БИЗНЕС - ПРОГНОЗИРОВАНИЕ

3.1 Теоретические сведения

Рассмотрим методы, которые используются для прогнозирования неопределенного будущего с целью помочь менеджерам в принятии наилучшего решения. Эти методы состоят в изучении и анализе накопленных данных с целью нахождения моделей, которые могут быть эффективно продолжены в будущее.

С развитием и ростом сложности аппарата прогнозирования, а также с появлением компьютеров, оснащенных соответствующим программным обеспечением, прогнозированию уделяется все больше и больше внимания. Сейчас каждый менеджер имеет реальную возможность использовать в прогнозировании очень сложный математический аппарат анализа данных, и знание этого аппарата является для него весьма существенным. По этой же причине менеджеры, использующие прогнозы в своей деятельности, должны понимать опасность выбора неадекватных методов прогнозирования, так как некорректные прогнозы могут привести к принятию неверных решений.

Прогнозы могут классифицироваться как долгосрочные и краткосрочные. Долгосрочные прогнозы необходимы для того, чтобы наметить основной курс предприятия на длительный период, поэтому именно на них акцентируется основное внимание менеджеров высшего звена. Краткосрочные прогнозы используются для разработки безотлагательных стратегий. Они чаще применяются менеджерами среднего и низшего звена для удовлетворения потребностей ближайшего будущего.

Как правило, под прогнозированием понимается процедура предсказания важных показателей для отдельных компаний или даже одного из подразделений компании. Примерами могут служить месячный объем продаж компании, объем продаж отдельных видов продукции для одного из магазинов компании или же количество пропущенных рабочих часов, которое приходится на одного работника фабрики. Как известно, прогнозирование продаж актуально практически для каждой компании. Качественный прогноз является первым шагом в решении множества бизнес-задач: оптимизация закупок, распределение ресурсов, минимизация кассовых разрывов,

бюджетирование. Для некоторых компаний прогнозирование стает жизненно важной задачей, которая может существенно способствовать финансовой стабильности и укреплению позиций на рынке.

В противоположность этому, наблюдается растущий интерес к прогнозированию важных параметров экономики всей страны. Например, правительство интересует прогноз уровня безработицы, роста национального продукта и значения основной учетной ставки. В частности, вся экономическая политика строится на планировании основных экономических показателей.

При выборе метода прогнозирования следует учитывать несколько факторов. Следует определить уровень детализации. Нужен ли прогноз определенных деталей (микро-прогноз)? Или же требуется прогноз будущего состояния всеобъемлющих или обобщенных факторов (макропрогноз)? Необходим ли прогноз в ближайшем будущем (краткосрочный прогноз) или в отдаленном будущем (долгосрочный прогноз)? И в какой степени являются приемлемыми качественные (оценочные) и количественные (оперирующие данными) методы прогнозирования?

Первым и, наверное, самым важным условием качественного прогнозирования является наличие достаточного количества исходных данных. На данном этапе уже много предприятий Украины не только внедрили учетные системы класса ERP и CRM в свою деятельность, но и накопили в них достаточный для анализа объем данных. Поэтому проблема количества исходных данных перестает быть критической. Очевидно, что это не снимает требования к качеству данных.

Так как аппарат прогнозирования оперирует данными, порожденными естественными событиями, определяют следующие пять этапов в процессе прогнозирования.

1. Сбор данных.
2. Редукция или уплотнение данных. Предварительная обработка данных, удаление аномальных значений и сглаживание шумов в данных.
3. Построение модели и ее оценка.
4. Экстраполяция выбранной модели (фактический прогноз).
5. Оценка полученного прогноза.

Этап 1, сбор данных, предполагает получение корректных данных и обязательную проверку того, что они верны. Этот этап часто является наиболее сомнительной частью всего процесса прогнозирования и в то же время наиболее сложен для проверки. Часто сбор и проверка данных сопровождается множеством различных проблем.

Этап 2, редукция или уплотнение данных, часто оказывается необходимым, так как для выполнения прогнозирования может быть собрано как слишком много исходных данных, так и слишком мало. Кроме того, данные могут иметь случайные аномальные выбросы, которые лучше убрать перед построением модели.

Этап 3, построение модели и ее оценка, состоит в подборе модели прогноза, наиболее соответствующей особенностям собранных данных в смысле минимизации ошибки прогноза.

Этап 4, экстраполяция выбранной модели, предусматривает фактическое получение требуемого прогноза. Часто для проверки точности получаемых результатов применяется прогнозирование на недавно прошедшие периоды, для которых исследуемые величины уже известны. Наблюдаемые ошибки затем определенным образом анализируются на этапе 5.

Этап 5, оценка точности и адекватности построенной модели и прогноза.

3.2 Компьютерные пакеты для решения задач прогнозирования

Таблица 3.1 – Классификация программных продуктов для расчетов прогнозов

Название инструмента	**Сфера применения**	**Реализуемые модели**	**Требуемая подготовка пользователя**	**Готовность к эксплуата-ции**
Microsoft Excel, OpenOffice.org	широкого назначения	алгоритмические, регрессионные	базовые знания статистики	требуется значительная доработка (реализация моделей)

Statistica, SPSS, E-views, Gretl	Исследовательская	широкий спектр регрессионных, нейросетевые	специальное математическое образование	коробочный продукт
Matlab	Исследоват., разработка приложений	алгоритмические, регрессионные, нейросетевые	специальное математическое образование	требуется программирование
ForecastPro, ForecastX Deductor	бизнес-прогнозирование	алгоритмические	не требуются глубокие знания	коробочный продукт
iLog, AnyLogic, iThink, Matlab Simulink, GPSS	разработка приложений, моделирование	имитационные	требуется специальное математическое образование	требуется программирование (под специфику области)

3.3 Временные ряды

Информационной базой для анализа экономических процессов являются динамические и временные ряды**.** Совокупность наблюдений некоторого явления (показателя), упорядоченная в зависимости от последовательности значений другого явления (признака), называют ***динамическим рядом***. Динамические ряды, у которых в качестве признака упорядочения используется время, называют ***временными***.

В экономике и бизнесе временные ряды – это очень распространенный тип данных. Во временном ряде содержится информация об особенностях и закономерностях протекания процесса, а статистический анализ позволяет выявить и использовать выявленные закономерности для оценки характеристик процесса в будущем, т.е. для прогнозирования.

Временной ряд – это набор чисел, привязанный к последовательным, обычно равноотстоящим моментам времени. Числа, составляющие временной ряд и получающиеся в результате наблюдения за ходом некоторого процесса, называются *уровнями* временного ряда или *элементами.* Под длиной временного ряда

понимают количество входящих в него уровней *n*. Временной ряд обычно обозначают *Y(t),* или y_t , где $t=1,2,\ldots,n$.

Временным рядом (ВР) $\{Y_t\}$ будем называть множество значений некоторой величины в последовательные моменты времени.

$$\{Y(t)\} = Y_1; Y_2; \ldots Y_{t-1}; Y_t; Y_{t+1}; \ldots$$

Прогнозирование временного ряда - вычисление величины его будущих значений либо характеристик, позволяющих определить эту величину, на основании анализа известных значений. Величина, подлежащая прогнозу, называется прогнозируемой величиной (ПВ).

При прогнозировании предполагается, что значение прогнозируемой величины зависит от каких-либо факторов, назовем их определяющими факторами, или признаками. Один из подходов к задаче прогнозирования основан на предположении зависимости ПВ от предыдущих значений ВР.

Пример графика временного ряда приведен на рис. 3.1.

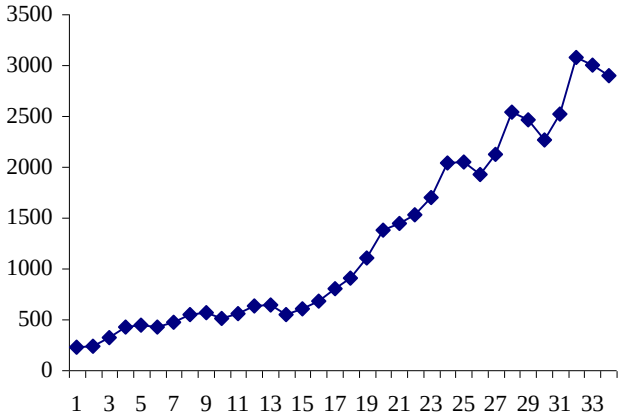


Рисунок 3.1 - График динамики временного ряда номинальный объем валового внутреннего продукта – квартальные данные

До недавнего времени (середины 80-х годов прошлого века) существовало несколько общепризнанных методов прогнозирования временных рядов:

- эконометрические;
- регрессионные ;
- методы Бокса-Дженкинса (ARIMA, ARMA).

Однако, начиная с конца 80-х годов, в научной литературе был опубликован ряд статей по нейросетевой тематике, в которых был приведен эффективный алгоритм обучения нейронных сетей и доказана возможность их использования для самого широкого круга задач. Одним из самых успешных приложений нейронных сетей было прогнозирование временных рядов. Причем самым массовым было

- прогнозирование на финансовых рынках;
- прогнозирование продаж.

В настоящее время можно с уверенностью сказать, что использование нейронных сетей при прогнозировании дает ощутимое преимущество по сравнению с более простыми статистическими методами.

3.3.1 Основные описательные статистики для временных рядов

Среднее и дисперсия временного ряда рассчитываются по формулам:

$$\bar{Y} = \sum_{t=1}^{T} Y_t, \quad s^2 = \frac{1}{T}\sum_{t=1}^{T} (Y_t - \bar{Y})^2.$$

Выборочная автоковариация *k*-го порядка вычисляется как

$$c_k = \frac{1}{T}\sum_{t=1}^{T-k} (Y_t - \bar{Y})(Y_{t+k} - \bar{Y})$$

Статистической оценкой автокорреляции *k*-го порядка для стационарных процессов является выборочный коэффициент автокорреляции: $r_k = c_k / c_0$. При анализе изменения величин c_k и r_k в зависимости от значения *k* обычно пользуются выборочными автоковариационной и автокорреляционной функциями, определяемыми как последовательности $\{c_k\}$ и $\{r_k\}$, соответственно. Выборочная автокорреляционная функция играет особую роль в

анализе стационарных временных рядов, поскольку может быть использована в качестве инструмента для распознавания типа процесса. При этом обычно анализируют график автокорреляционной функции, называемый коррелограммой.

Стационарным процессом называется такой случайный процесс, вероятностные свойства которого с течением времени не изменяются. Он протекает в приблизительно однородных условиях и имеет вид непрерывных случайных колебаний вокруг некоторого среднего значения. Причем ни средняя амплитуда, ни его частота не обнаруживают с течением времени существенных изменений.

3.3.2 Аддитивные модели временных рядов

В общем случае каждый уровень временного можно представить как функцию четырех компонент: $f(t)$, $S(t)$, $U(t)$, $\varepsilon(t)$, отражающих закономерность и случайность развития. Где

$f(t)$ – тренд (долговременная тенденция) развития;
$S(t)$ – сезонная компонента;
$U(t)$ –циклическая компонента;
$\varepsilon(t)$– остаточная компонента.

В модели временного ряда принято выделять две основные составляющие: детерминированную (систематическую) и случайную. Под детерминированной составляющей временного ряда $y_1, y_2, \ldots, y_n$ понимают числовую последовательность, элементы которой вычисляются по определенному правилу как функция времени t. Исключив детерминированную составляющую из данных, мы получим колеблющийся вокруг нуля ряд, который может в одном предельном случае представлять случайные скачки, а в другом – плавное колебательное движение.

Детерминированная составляющая может содержать следующие структурные компоненты:

1) *Тренд,* или *тенденция* $f(t)$, представляет собой устойчивую закономерность, наблюдаемую в течение длительного периода времени. Обычно тренд (тенденция) описывается с помощью той или иной неслучайной функции $f_{тр}(t)$ (аргументом которой является

время), как правило, монотонной. Эту функцию называют функцией тренда, или просто – трендом.

2) *Сезонная компонента s(t)* связана с наличием факторов, действующих с заранее известной периодичностью. Это регулярные колебания, которые носят периодический или близкий к нему характер и заканчиваются в течение года. Типичные примеры сезонного эффекта: изменение загруженности автотрассы по временам года, пик продаж товаров для школьников в конце августа – начале сентября. Спрос на пластические операции сезонный: в осенне-зимний период обращений больше. Типичным примером являются сильные колебания объема товарно-материальных запасов в сезонных отраслях Сезонная компонента со временем может меняться, либо иметь плавающий характер.

3) *Циклическая компонента* $U(t)$ – неслучайная функция, описывающая длительные периоды (более одного года) относительного подъема и спада и состоящая из циклов переменной длительности и амплитуды. Примером циклической (конъюнктурной) компоненты являются волны Кондратьева, демографические «ямы» и т.п. Подобная компонента весьма характерна для рядов макроэкономических показателей. Здесь циклические изменения обусловлены взаимодействием спроса и предложения, а также наложением таких факторов, как истощение ресурсов, погодные условия, изменения в налоговой политике и т.п. Отметим, что циклическую компоненту крайне трудно идентифицировать формальными методами, исходя только из данных изучаемого ряда.

4) Случайная компонента $\varepsilon(t)$ - это составная часть временного ряда, оставшаяся после выделения систематических компонент. Она отражает воздействие многочисленных факторов случайного характера и представляет собой случайную, нерегулярную компоненту. Она является обязательной составной частью любого временного ряда в экономике, так как случайные отклонения неизбежно сопутствуют любому экономическому явлению. Если систематические компоненты временного ряда определены правильно, то остающаяся после выделения из временного ряда этих компонент так называемая остаточная последовательность (ряд остатков) будет случайной компонентой ряда.

В анализе случайного компонента экономических временных рядов важную роль играет сравнение случайной величины ε_t с хорошо изученной формой случайных процессов - стационарными случайными процессами.

В зависимости от вида связи между этими компонентами может быть построена либо аддитивная модель:

$$Y(t) = f(t) + S(t) + U(t) + \varepsilon(t); \qquad (3.1)$$

либо мультипликативная модель:

$$Y(t) = f(t) \cdot S(t) \cdot U(t) + \varepsilon(t) \qquad (3.2)$$

В процессе формирования значений временных рядов не всегда участвуют все четыре компоненты. Однако во всех случаях предполагается наличие ***случайной составляющей.***

Тренды. Проводя разложение ряда на компоненты, мы, как правило, подразумеваем под трендом изменение среднего уровня переменной, то есть тренд среднего.

В рамках анализа тренда среднего чаще всего используют ***полиномиальный тренд:***

$$Y_t = a_0 + a_1 t + \ldots + a_p t^p \qquad (3.3)$$

Для $p = 1$ имеем линейный тренд.

3.3.3 Авторегрессионные модели временных рядов

Обозначаются ***AR(p) -авторегрессионая модель порядка p.*** Модель имеет вид:

$$Y_t = f_0 + f_1 \cdot Y_{t-1} + f_2 \cdot Y_{t-2} + \ldots + f_p \cdot Y_{t-p} + E_t \qquad (3.4)$$

где Y_t - зависимая переменная в момент времени t.

$f_0, f_1, \ldots, f_p$ - оцениваемые параметры. E_t - ошибка от влияния переменных, которые не учитываются в данной модели.

Задача заключается в том, чтобы определить $f_0, f_1, \ldots, f_p$. Их можно оценить различными способами. Один из наиболее простых способов - посчитать их методом наименьших квадратов.

Термин ***авторегрессия*** для обозначения модели (3.4) используется потому, что она фактически представляет собой модель регрессии, в которой регрессорами служат лаги изучаемого ряда Y_t. По определению авторегрессии ошибки E_t являются белым шумом и некоррелированы с лагами Y_t. Таким образом, выполнены все основные предположения регрессионного анализа: ошибки имеют нулевое математическое ожидание, некоррелированы с регрессорами, не автокоррелированы и гомоскедастичны. Следовательно, модель (3.4) можно оценивать с помощью обычного метода наименьших квадратов. Отметим, что при таком оценивании p начальных наблюдений теряются.

После построения любой модели временного ряда, прежде, чем прогнозировать по этой модели, нужно убедиться в ее адекватности, т.е. убедиться, что остатки E_t некоррелированы между собой.

Критерий Дарбина-Уотсона является наиболее распространенным критерием для проверки корреляции внутри ряда. Если величина

$$D = \frac{\sum_{i=2}^{n} (e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2}$$, где e_i - расхождение между фактическими и расчетными уровнями, имеет значение, близкое к 2, то можно считать модель достаточно адекватной. Когда адекватная модель найдена, можно делать прогнозы на один или несколько периодов вперед.

3.3.4 Нейросетевые модели прогнозирования

В настоящее время самым перспективным количественным методом прогнозирования является использование нейронных сетей. Можно назвать много преимуществ нейронных сетей над остальными алгоритмами, ниже приведены два основных.

Построение нейросетевой модели происходит адаптивно во время обучения, без участия эксперта. При этом нейронной сети предъявляются примеры из базы данных и она сама подстраивается под эти данные.

Искусственный нейрон и нейронная сеть

Несмотря на большое разнообразие вариантов нейронных сетей, все они имеют общие черты. Так, все они, так же, как и мозг человека, состоят из большого числа связанных между собой однотипных элементов – *нейронов*, которые имитируют нейроны головного мозга. На рис. 3.2 показана схема нейрона.

Из рисунка видно, что искусственный нейрон, так же, как и живой, состоит из синапсов, связывающих входы нейрона с ядром; ядра нейрона, которое осуществляет обработку входных сигналов и аксона, который связывает нейрон с нейронами следующего слоя. Каждый синапс имеет вес, который определяет, насколько соответствующий вход нейрона влияет на его состояние. Состояние нейрона определяется по формуле

$$S = \sum_{i=1}^{n} x_i w_i \qquad (3.5)$$

где n – число входов нейрона; x_i – значение i-го входа нейрона;
w_i – вес i-го синапса

Рисунок 3.2 – Схема нейрона

Затем определяется значение аксона нейрона по формуле

$$Y = f(S) \qquad (3.6)$$

где f – некоторая функция, которая называется *активационной*. Наиболее часто в качестве активационной функции используется так называемый *сигмоид*, который имеет следующий вид:

$$f(x) = \frac{1}{1+e^{-\alpha \cdot x}} \qquad (3.7)$$

При уменьшении параметра α сигмоид становится более пологим, вырождаясь в горизонтальную линию на уровне 0,5 при α =0. При увеличении α сигмоид все больше приближается к функции единичного скачка.

Сегодня существует большое число различных конфигураций нейронных сетей с различными принципами функционирования, которые ориентированы на решение самых разных задач. В качестве примера рассмотрим многослойную полносвязанную нейронную сеть прямого распространения (рис.3.3), которая широко используется для поиска закономерностей и классификации образов. Полносвязанной нейронной сетью называется многослойная структура, в которой каждый нейрон произвольного слоя связан со всеми нейронами предыдущего слоя, а в случае первого слоя — со всеми входами нейронной сети. Прямое распространение сигнала означает, что такая нейронная сеть не содержит петель.

Пусть N общее число нейронов в сети;

n_j - число входов в j - й нейрон;

x_{ji} - значение силы импульса i - го входа в j - й нейрон;

w_{ji} - вес импульса i - го входа в j - й нейрон;

$$S_j = \sum_{i \in n_j} w_{ji} \cdot x_{ji} \qquad (3.8)$$

$y_j = f(S_j)$ - выход j - ого нейрона.

Определим функцию ошибки для нейронной сети:

$$E(W) = \frac{1}{2}\sum_{j=1}^{p}(y_j - d_j)^2 , \qquad (3.9)$$

Где d_j - целевое значение выхода j - ого нейрона;

p - число нейронов в выходном слое.

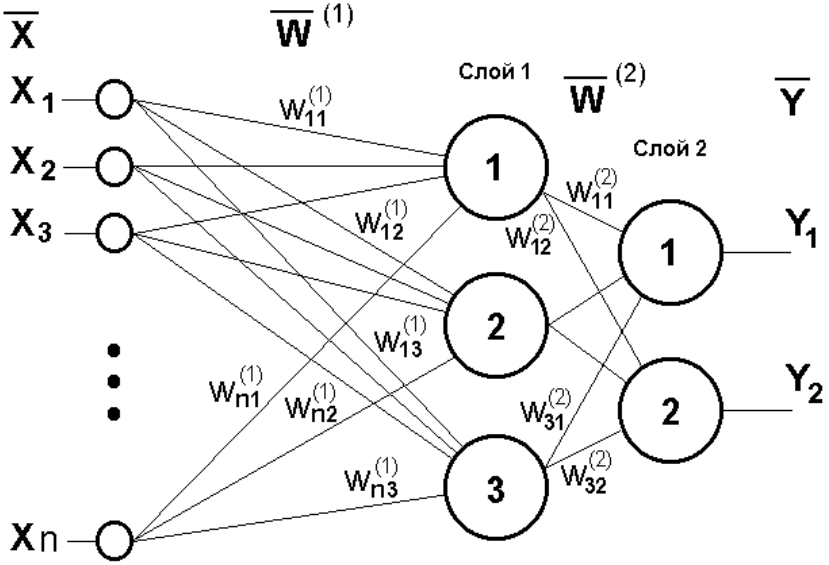


Рисунок 3.3 – Двухслойная нейронная сеть (перцептрон)

Обучить нейросеть означает найти весовые коэффициенты w_{ji} так, чтобы функция ошибки была минимальной.

Обучение

Способность к обучению является основным свойством мозга. Для искусственных нейронных сетей под обучением понимается процесс настройки архитектуры сети (структуры связей между нейронами) и весов синаптических связей (влияющих на сигналы коэффициентов) для эффективного решения поставленной задачи. Обычно обучение нейронной сети осуществляется на некоторой выборке. По мере процесса обучения, который происходит по некоторому алгоритму, сеть должна все лучше и лучше (правильнее) реагировать на входные сигналы.

Выделяют три парадигмы обучения: с учителем, самообучение и смешанная. В первом способе известны правильные ответы к каждому

входному примеру, а веса подстраиваются так, чтобы минимизировать ошибку. Обучение без учителя позволяет распределить образцы по категориям за счет раскрытия внутренней структуры и природы данных. При смешанном обучении комбинируются два вышеизложенных подхода.

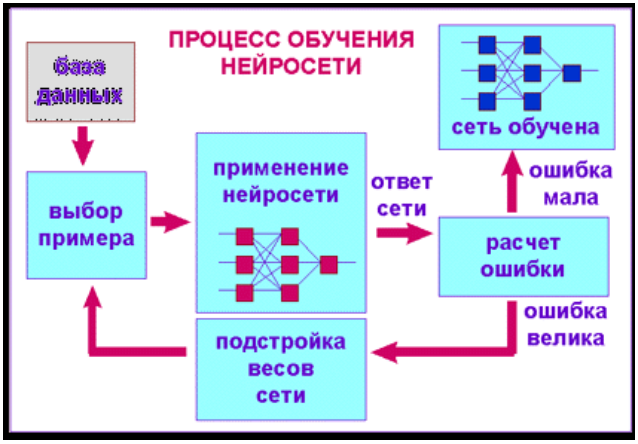


Рисунок 3.4 – Схема обучения нейронной сети

Подготовка данных для обучения нейронной сети

Пусть у нас имеется база данных, содержащая значения курса за последние 300 дней. Простейший вариант в данном случае - попытаться построить прогноз завтрашней цены на основе курсов за последние несколько дней. Понятно, что прогнозирующая нейронная сеть должна иметь всего один выход и столько входов, сколько предыдущих значений мы хотим использовать для прогноза - например, 4 последних значения. Составить обучающий пример очень просто - входными значениями нейронной сети будут курсы за 4 последовательных дня, а желаемым выходом нейронной сети - известный нам курс в следующий день за этими четырьмя.

Предобработка данных

На практике большинство прогнозируемых временных рядов порождаются сложными динамическими системами, с множеством степеней свободы. Кроме того, в самом временном ряде может присутствовать случайная составляющая. Поэтому необходимо

выполнить предобработку данных, что позволяет уменьшить ошибку прогнозирования.

3.4 Пример прогнозирования с помощью линейной регрессии

Программа Deductor содержит механизмы импорта, обработки, визуализации и экспорта данных для быстрого и эффективного анализа и прогнозирования.

В документации к Deductor Studio приведен пример построения законченного решения по прогнозированию объемов продаж товаров на три месяца вперед.

Рассмотрим группу инструментов предобработки данных, которая приводит исходные, "сырые", данные к виду, пригодному для анализа и обработки. Затем рассмотрим механизмы преобразования данных, которые модифицируют данные на основе настроек аналитика. И, наконец, рассмотрим алгоритмы анализа данных, позволяющие находить зависимости одних факторов от других, кластеризовать данные, обнаружить сезонность во временных рядах, а также построить модель прогноза и получить желаемый результат (провести эксперимент, спрогнозировать временной ряд).

3.4.1 Импорт данных из файла

Импорт осуществляется путем вызова Мастера импорта на панели "Сценарии". После запуска Мастера импорта укажем тип импорта "Текстовый файл с разделителями" и перейдем к настройке импорта. Отметим имя файла, из которого необходимо получить данные – Trade.txt. В окне просмотра выбранного файла можно увидеть содержание данного файла (рис. 3.4).

Далее перейдем к настройке параметров импорта. На этой странице Мастера предоставляется возможность указать, с какой строки следует начать импорт, отметить то, что первая строка является заголовком, что является символом-разделителем столбцов, а также ограничитель строк, разделитель целой и дробной частей вещественного числа (рис. 3.5). В данном случае параметры по умолчанию на этой странице Мастера установлены правильно, а именно начать импорт с первой строки, первая строка является заголовком, разделителем между столбцами является знак табуляции, разделителем целой и дробной частей является запятая.

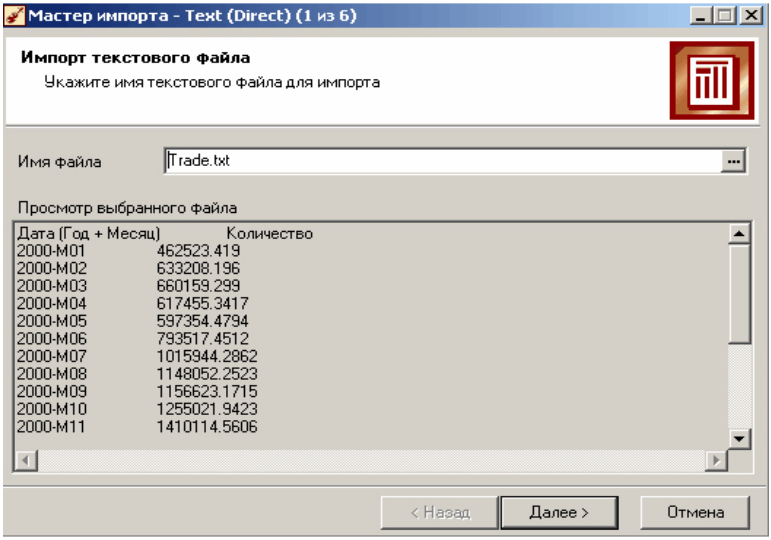


Рисунок 3.4 – Второе окно мастера импорта

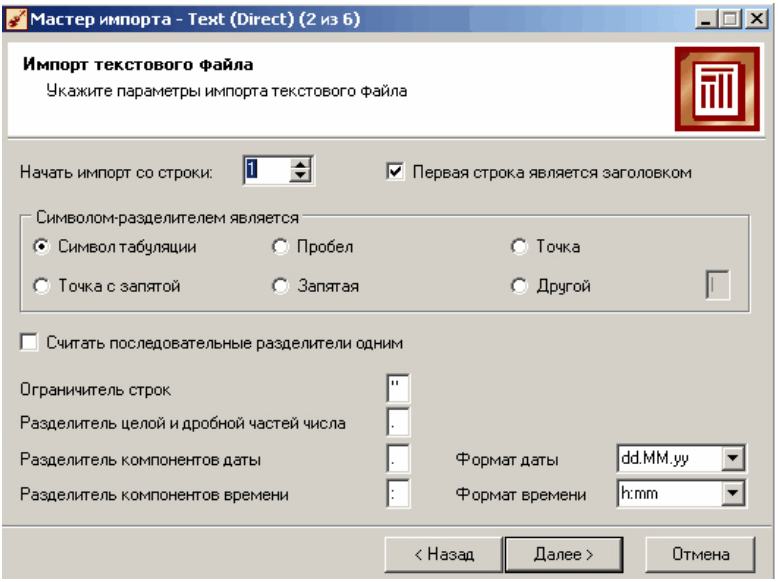


Рисунок 3.5 – Третье окно мастера импорта

3.4.2 Настройка параметров столбцов

На этом шаге Мастера предоставляется возможность настроить параметры каждого поля: имя, метку, размер, тип данных, вид данных и назначение. Некоторые свойства (например, тип данных) можно задавать сразу для группы полей. Вид данных определяет, конечный ли это набор (дискретные) или бесконечный (непрерывные). Назначение столбцов выявляет характер их использования в алгоритмах обработки (при импорте можно оставить значение по умолчанию).

Указав параметры столбцов, запустим процесс импорта, нажав на кнопку "Пуск". После импорта данных на следующем шаге Мастера необходимо выбрать способ отображения данных. В данном случае самым информативным является диаграмма, поэтому выберем ее. После настройки отображения необходимых полей получаем график временного ряда исходных данных на рис. 3.7.

Перейдем к предварительной обработке загруженных данных. Все шаги по подготовке и обработке данных представлены в виде сценария на рисунке 3.8.

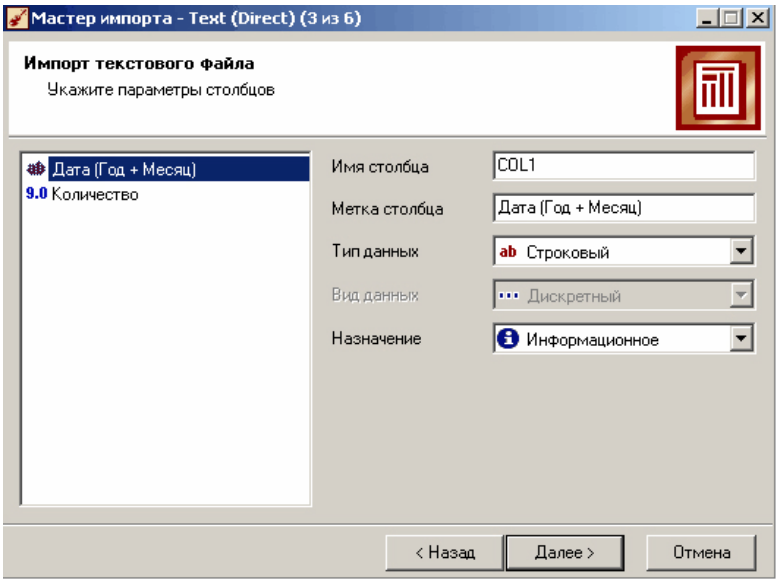


Рисунок 3.6 – Настройка параметров столбцов

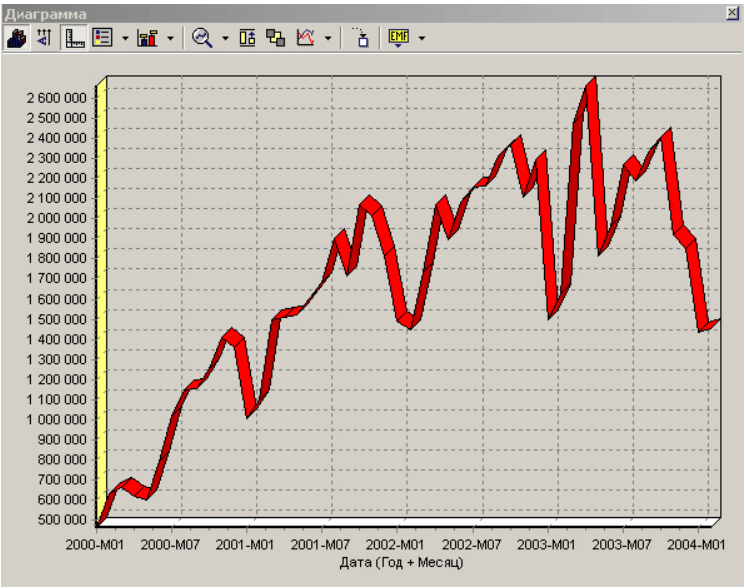


Рисунок 3.7 – Временной ряд исходных данных

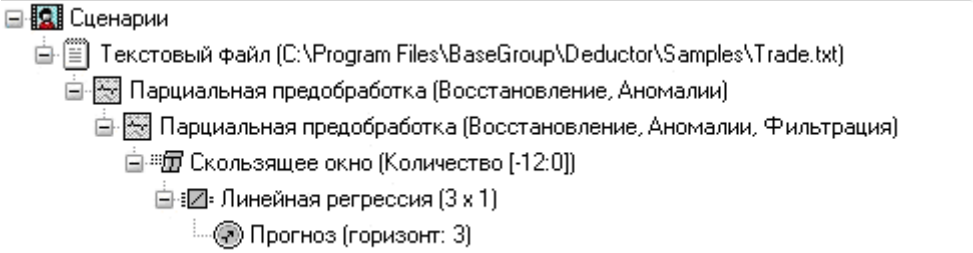


Рисунок 3.8 – Сценарий прогнозирования продаж

Парциальная предобработка служит для восстановления пропущенных данных, редактирования аномальных значений и спектральной обработки данных (например, сглаживания данных). Именно эти операции часто проводятся в первую очередь над данными.

3.4.3 Расчет автокорреляции столбцов

Важным фактором для анализа временного ряда и прогноза является определение сезонности. В **Deductor Studio** инструментом,

предназначенным для изучения сезонности, является автокорреляция. Вообще корреляция подразумевает под собой зависимость значения одной величины от значения другой. Если их корреляция равна единице, то величины прямо зависимы друг от друга, если нулю, то нет, если минус единица, то зависимость обратная. Нахождение линейной автокорреляционной зависимости применяется для определения периодичности (сезонности) при обработке временных рядов.

Как видно, не каждый аналитик сможет судить о сезонности по этим данным, поэтому необходимо воспользоваться автокорреляцией. Для этого откроем Мастер обработки, выберем в качестве обработки автокорреляцию и перейдем на второй шаг Мастера. В нем необходимо настроить параметры столбцов. Укажем поле "Дата (Год + Месяц)" неиспользуемым, а поле "Количество" используемым (ведь необходимо определить сезонность количества продаж). Предположим, что сезонность, если она имеет место, не больше года. В связи с этим зададим Количество отсчетов равным 15 (тогда будет искаться зависимость от месяца назад, двух, ..., пятнадцати месяцев назад). Количество отсчетов ставится больше 12 (хотя мы ищем наличие именно готовой сезонности, т.е. 12 месяцев) для того, чтобы убедиться, что на 12 месяцев приходится пик коэффициента автокорреляции, а далее следует его спад.

Также должен стоять флажок "Включить поле отсчетов набор данных". Он необходим для более удобной интерпретации автокорреляционного анализа.

Перейдем на следующий шаг Мастера и запустим процесс обработки.

По окончанию результаты удобно анализировать как в виде таблицы, так и в виде диаграммы. После обработки были получены два столбца – "Лаг" (благодаря установленному флажку в Мастере) и "Количество" - результат автокорреляции.

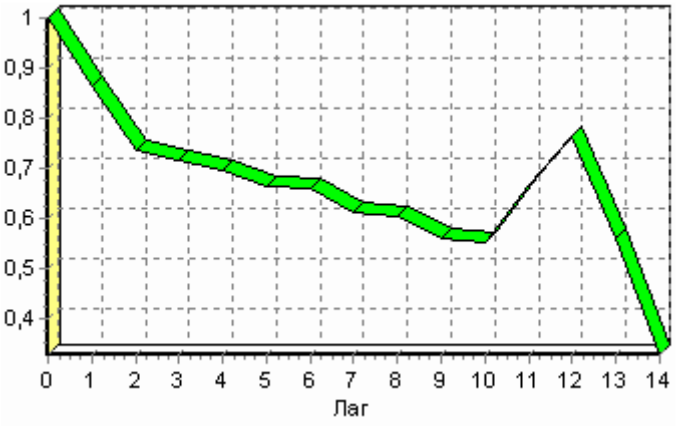


Рисунок 3.9 – Функция автокорреляции временного ряда

Видно, что вначале корреляция равна единице – так как значение зависит само от себя. Далее зависимость убывает, и затем виден пик зависимости от данных 12 месяцев назад. Это как раз и говорит о наличии годовой сезонности.

3.4.4 Удаление аномалий

Присутствие аномалий при построении моделей оказывает на них большое влияние, ухудшая качество результата. Как видно из диаграммы, выбросы ухудшают статистическую картину распределения данных. Воспользуемся Мастером обработки и выберем парциальную обработку.

В Мастере парциальной предобработки на втором шаге выбираем поле "Количество" и указываем ему тип обработки "Редактирование аномальных значений", степень подавления "Большая". Так как больше никаких действий над данными не планировалось, то переходим на шаг запуска процесса обработки и нажимаем "Пуск".

После выполнения процесса обработки на диаграмме (рис. 3.11) видно, что выбросы уменьшились, и стала проясняться реальная картина продаж.

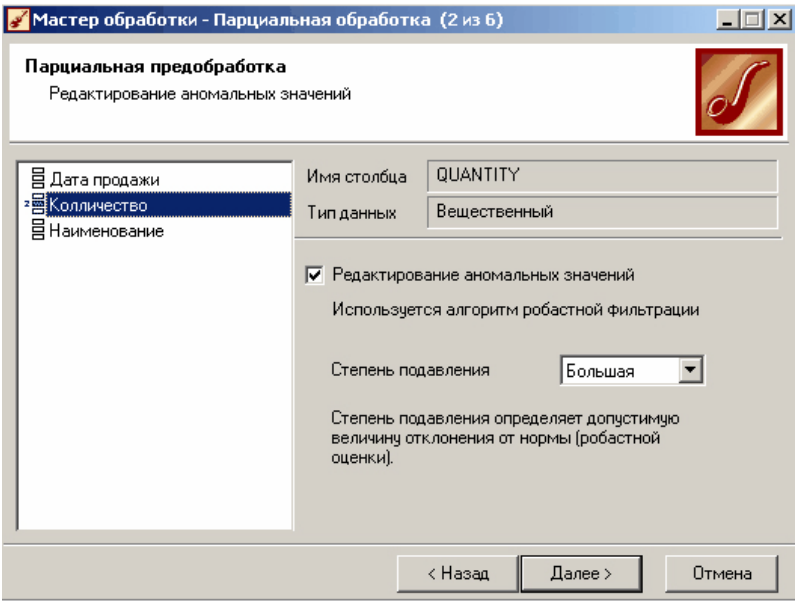


Рисунок 3.10 – Настройка параметров парциальной предобработки

3.4.5 Сглаживание данных – удаление шумов

Сглаживание данных применяется для удаления шумов из исходного набора (что будет продемонстрировано позднее), а также для выделения тенденции, трудно обнаруживаемой в исходном наборе. Платформа Deductor Studio предлагает несколько видов спектральной обработки: сглаживание данных путем указания полосы пропускания, вычитание шума путем указания степени вычитания шума и вейвлета преобразования путем указания глубины разложения и порядка вейвлета.

Сгладим данные при помощи парциальной обработки.

В Мастере парциальной предобработки на третьем шаге выбираем поле "Количество" и указываем ему тип обработки "Вычитание шума", степень подавления "Большая". Переходим на шаг запуска процесса обработки и нажимаем "Пуск".

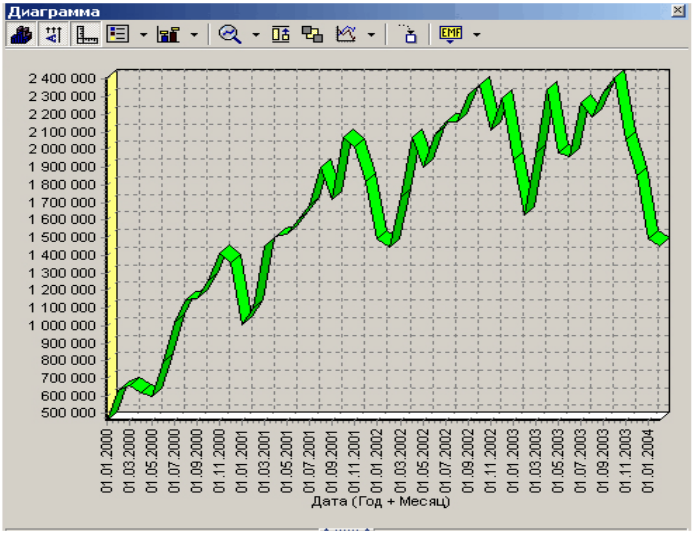


Рисунок 3.11 – Временной ряд после удаления аномалий

Как видно из диаграммы рис. 3.12 данные стали более сглаженными и могут служить для дальнейшей обработки. Взглянув на данные легко понять общую тенденцию.

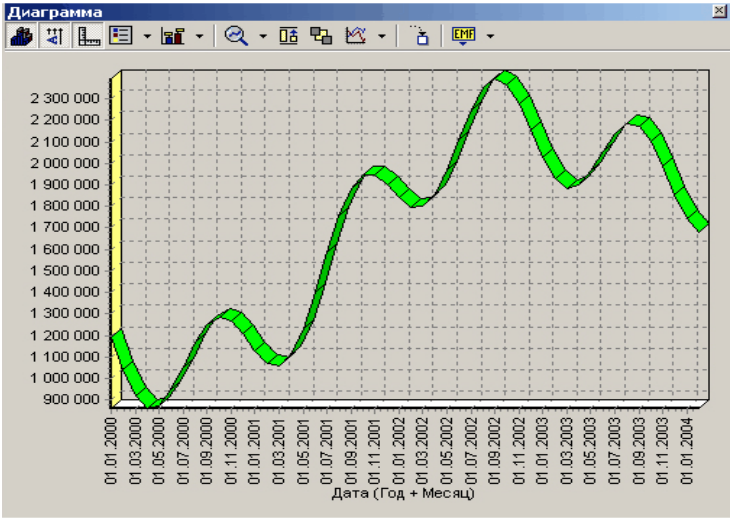


Рисунок 3.12 – Временной ряд после удаления шумов

3.4.6 Преобразование данных к скользящему окну

Когда требуется прогнозировать временной ряд, тем более, если налицо его периодичность (сезонность), то лучшего результата можно добиться, учитывая значения факторов не только в данный момент времени, но и, например, за аналогичный период прошлого года. Такую возможность можно получить после трансформации данных к скользящему окну. То есть, например, при сезонности продаж с периодом 12 месяцев, для прогнозирования количества продаж на месяц вперед можно в качестве входного фактора указать не только значение количества продаж за предыдущий месяц, но и за 12 месяцев назад.

Обработка создает новые столбцы путем сдвига данных исходного столбца вниз и вверх (глубина погружения и горизонт прогноза).

У аналитика имеются данные о месячном количестве проданного товара за несколько лет. Ему необходимо, основываясь на этих данных, сказать, какое количество товара будет продано через неделю и через две.

Запустим Мастер обработки, выберем в качестве обработчика скользящее окно и перейдем на следующий шаг.

Можно использовать обработчик "Автокорреляция" и убедиться в наличии годовой сезонности. В связи с этим строить прогноз на месяц вперед можно, основываясь на данных за 1, 2, 11 и 12 месяцев назад. Поэтому необходимо, назначив поле "Количество" используемым, выбрать глубину погружения 12. Тогда данные трансформируются к скользящему окну так, что аналитику будут доступны все требуемые факторы для построения прогноза.

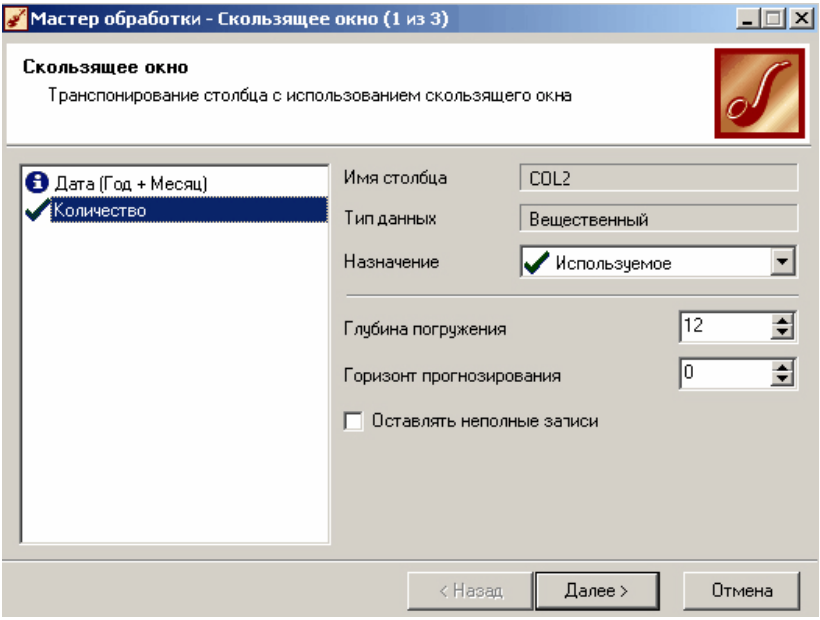


Рисунок 3.13 – Преобразование данных к скользящему окну

Просмотреть полученные данные можно в виде таблицы.

Дата (Год + Месяц)	Количество-12	Количество-11	Количество-10	Количество-9	Количество-8	Количество-7
2001-M01	1194372,7	1038792,3	919614,2	861513,6	873411,2	946129,4
2001-M02	1038792,3	919614,2	861513,6	873411,2	946129,4	1055162,7
2001-M03	919614,2	861513,6	873411,2	946129,4	1055162,7	1167771,5
2001-M04	861513,6	873411,2	946129,4	1055162,7	1167771,5	1252378,4
2001-M05	873411,2	946129,4	1055162,7	1167771,5	1252378,4	1287590,9
2001-M06	946129,4	1055162,7	1167771,5	1252378,4	1287590,9	1268284,4
2001-M07	1055162,7	1167771,5	1252378,4	1287590,9	1268284,4	1207012,7
2001-M08	1167771,5	1252378,4	1287590,9	1268284,4	1207012,7	1130347,5
2001-M09	1252378,4	1287590,9	1268284,4	1207012,7	1130347,5	1071189,9
2001-M10	1287590,9	1268284,4	1207012,7	1130347,5	1071189,9	1059244,3
2001-M11	1268284,4	1207012,7	1130347,5	1071189,9	1059244,3	1112368,9
2001-M12	1207012,7	1130347,5	1071189,9	1059244,3	1112368,9	1231268,6
2002-M01	1130347,5	1071189,9	1059244,3	1112368,9	1231268,6	1399044,9
2002-M02	1071189,9	1059244,3	1112368,9	1231268,6	1399044,9	1585735,8
2002-M03	1059244,3	1112368,9	1231268,6	1399044,9	1585735,8	1756560,3

Рисунок 3.14 – Данные, подготовленные для регрессионного анализа

Как видно, теперь в качестве входных факторов можно использовать "Количество-12", "Количество-11" - данные по количеству 12 и 11 месяцев назад (относительно прогнозируемого месяца) и остальные необходимые факторы. В качестве результата прогноза будет указан столбец "Количество".

3.4.7 Прогнозирование с помощью линейной регрессии

Линейная регрессия необходима тогда, когда предполагается, что зависимость между входными факторами и результатом линейная. Достоинством ее можно назвать быстроту обработки входных данных и простоту интерпретации полученных результатов.

Обучение линейной регрессии

Для построения линейной регрессии необходимо запустить Мастер обработки и выбрать в качестве обработки данных Линейную регрессию.

На первом шаге задаем назначение исходных столбцов. Предположим, что на прогноз влияет информация за 2 прошлых месяца и за два месяца год назад, тогда укажем входными столбцами поля: "Количество-12", "Количество-11", "Количество-2", и "Количество-1". В качестве выходного поля укажем столбец "Количество". Остальные поля помечаем как информационные или неиспользуемые.

На следующем шаге происходит настройка обучающего и тестового множеств, способ разложения исходного множества данных.

Третий шаг установки позволяет осуществить ограничение диапазона входных значений. Данный шаг оставим без изменений. При нажатии на кнопку "Далее" появляется окно запуска процесса обучения. В процессе выполнения видно, какая часть распознана на этапе обучения и теста.

После выполнения процесса выберем в качестве способа отображения диаграмму рассеяния и отображение результатов в виде диаграммы. Как видно из диаграммы рассеяния, обучение прошло с хорошей точностью.

Выходное поле: Количество

Атрибут	Коэффициент
9.0 <Константа>	1,3816E5
9.0 Количество-12	-0,45587
9.0 Количество-11	0,46057
9.0 Количество-2	-0,60455
9.0 Количество-1	1,524

Рисунок 3.15 – Коэффициенты полученной регрессионной модели

На рис. 3.15 приведены коэффициенты полученной модели. Запишем саму модель, обозначив Y_t - объем продаж в момент времени t:

$$Y_t = 1{,}382 \cdot 10^5 + 1{,}524 \cdot Y_{t-1} - 0{,}6045 \cdot Y_{t-2} + 0{,}4606 \cdot Y_{t-11} - 0{,}4559 \cdot Y_{t-12}$$

Возьмем теперь в качестве t - момент времени, следеющий за последним известным значением. Используя приведенную модель, можно будет вычислить объем продаж на месяц вперед. Повторив эту процедуру, вычислим объем продаж еще на месяц вперед и т.д.

В программе Deductor прогноз вычисляется автоматически.

Прогнозирование

Теперь для построения прогноза запустим Мастера обработки, в котором выберем прогнозирование. На первом шаге обработчика происходит настройка связи столбцов для прогнозирования. Укажем связь между столбцами и горизонт прогноза равный 3 (рис. 3.16).

На следующем шаге задаются параметры визуализации. Для данного примера выбираем отображение результатов в виде диаграммы прогноза. Теперь аналитик может дать прогноз о продажах, основываясь на модели, построенной с помощью линейной регрессии (рис. 3.17).

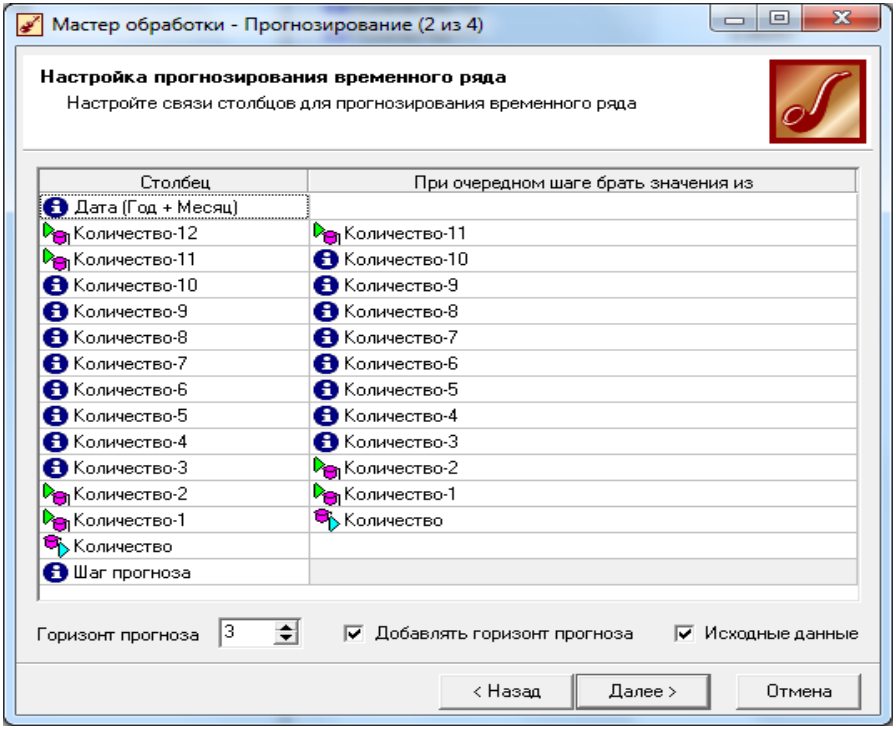


Рисунок 3.16 – Настройка параметров прогнозирования

Выводы

Данный пример показал целесообразность применения линейного регрессионного анализа для прогнозирования линейных зависимостей.

Простота настроек и быстрота построения модели иногда бывают необходимы. Аналитику достаточно указать входные столбцы - факторы, выходные – результат, указать способ разбиения данных на тестовое и обучающее множество и запустить процесс обучения. Причем после этого будут доступны все механизмы визуализации и анализа данных, позволяющие построить прогноз, провести эксперимент по принципу "Что-если", исследовать зависимость результата от значений входных факторов, оценить качество построенной модели по диаграмме рассеяния. Также по результатам работы этого алгоритма можно подтвердить или опровергнуть гипотезу о наличии линейной зависимости.

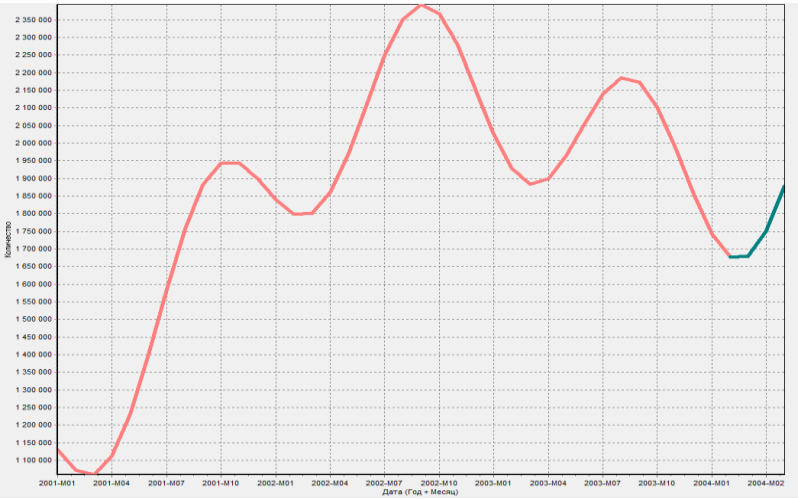


Рисунок 3.17 – Прогноз продаж на три периода вперед

3.5 Прогнозирование с помощью нейронных сетей

Прогнозирование появляется в списке Мастера обработки только после построения какой-либо модели прогноза: нейросети, линейной регрессии и т.д. Прогнозировать на несколько шагов вперед имеет смысл только временной ряд (к примеру, если есть данные по недельным суммам продаж за определенный период, можно спрогнозировать сумму продаж на две недели вперед). Поскольку при построении модели прогноза необходимо учитывать много факторов (зависимость результата от данных день, два, три, четыре назад), то методика имеет свои особенности. Покажем ее на примере.

3.5.1 Исходные данные

У аналитика имеются данные о помесячном количестве проданного товара за несколько лет. Ему необходимо, основываясь на этих данных, определить, какое Количество товара будет продано через месяц и через два.

Исходные данные по продажам находятся в файле "Trade.txt", известному по предыдущему примеру. Выполним импорт данных из файла, не забыв указать в Мастере, чтобы в качестве разделителя дробной и целой частей была точка, а не запятая.

3.5.2 Удаление аномалий и сглаживание

После импорта данных воспользуемся диаграммой для их просмотра. Как и в предыдущем пункте, перед прогнозированием необходимо удалить аномалии и сгладить данные. Сделаем это при помощи парциальной обработки как описано выше. Видно, что данные сгладились, аномалии и шумы исчезли. Также видна тенденция.

Теперь необходимо трансформировать данные к скользящему окну. Выбраем глубину погружения 12. Теперь в качестве входных факторов можно использовать "Количество-12", "Количество-11" - данные по количеству 12 и 11 месяцев назад (относительно прогнозируемого месяца), а также "Количество-2" и "Количество-1" - данные за 2 предыдущих месяца. В качестве выходного поля укажем столбец "Количество".

3.5.3 Обучение нейросети

Перейдем непосредственно к самому построению модели прогноза. Откроем Мастер обработки и выберем в нем нейронную сеть. На втором шаге Мастера согласно с принятым ранее решением установим в качестве входных поля "Количество - 12", "Количество - 11", "Количество - 2" и "Количество - 1", а в качестве выходного - "Количество". Остальные поля сделаем информационными.

На следующем шаге укажем разбиение тестового и обучающего множеств. Перейдем к следующему шагу, на котором отметим необходимое количество слоев и нейронов в нейросети (рис. 3.19).

Перейдя далее, выберем алгоритм обучения нейросети - RPROP.

После построения модели для просмотра качества обучения представим полученные данные в виде диаграммы и диаграммы рассеяния.

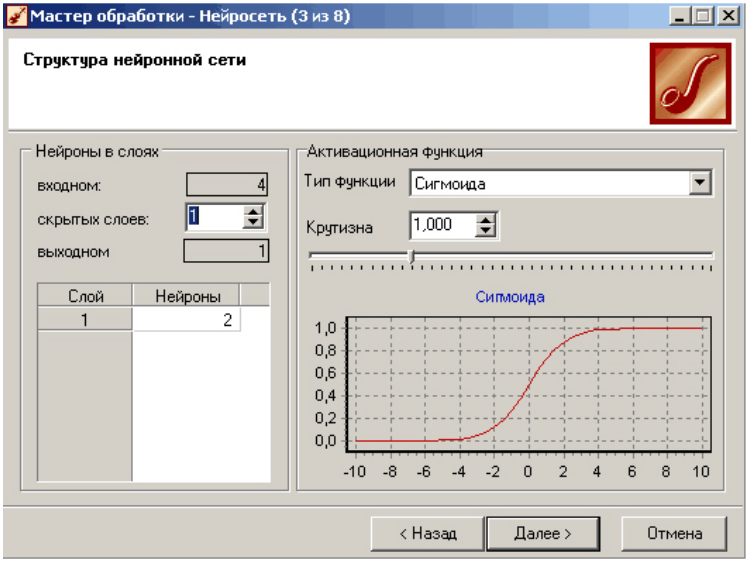


Рисунок 3.19 – Задание структуры нейронной сети

В Мастере настройки диаграммы выберем для отображения поля "Количество" и "Количество_OUT" - реальное и спрогнозированное значение. Результатом будет два графика (рис. 3.20).

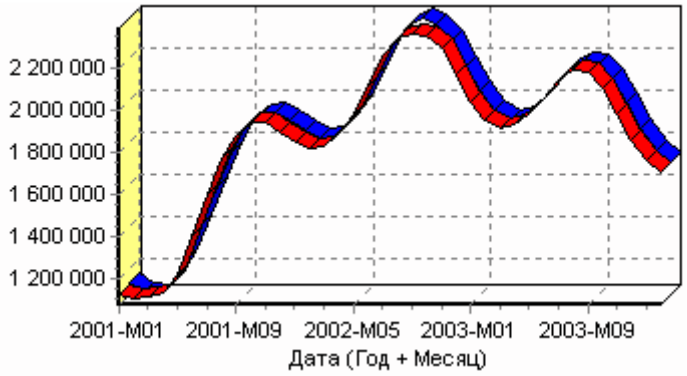


Рисунок 3.20 – Исходный и аппроксимированный нейросетью графики объемов продаж

3.5.4 Построение прогноза

Нейросеть обучена, осталось получить требуемый прогноз. Для этого открываем Мастер обработки и выбираем появившийся теперь обработчик "Прогнозирование".

На втором шаге Мастера предлагается настроить связи столбцов для прогнозирования временного ряда: откуда брать данные для столбца при очередном шаге прогноза. Мастер сам верно настроил все переходы, поэтому остается только указать горизонт прогноза (на сколько вперед будем прогнозировать) равный трем, а также для наглядности следует добавить к прогнозу исходные данные, установив в Мастере соответствующий флажок.

3.5.5 Результат

После этого необходимо в качестве визуализатора выбрать "Диаграмму прогноза", которая появляется только после прогнозирования временного ряда.

В Мастере настройки столбцов диаграммы прогноза надо указать в качестве отображаемого столбец "Количество", а в качестве подписей по оси X указать столбец "Шаг прогноза".

Теперь аналитик может дать ответ на вопрос, какое Количество товаров будет продано в следующем месяце и даже два месяца спустя.

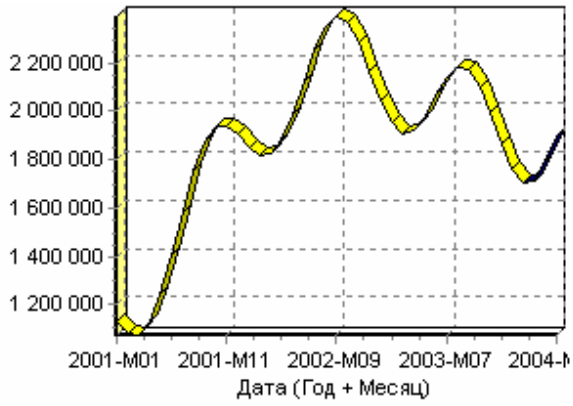


Рисунок 3.21 – Диаграмма прогноза временного ряда

3.5.6 Выводы

Данный пример показал, как с помощью **Deductor Studio** прогнозировать временной ряд.

При решении задачи были применены механизмы очистки данных от шумов, аномалий, которые обеспечили качество построения модели прогноза далее и соответственно достоверный результат самого прогнозирования количества продаж на три месяца вперед. Также был продемонстрирован принцип прогнозирования временного ряда – импорт, выявление сезонности, очистка, сглаживание, построение модели прогноза и собственно построение прогноза временного ряда.

Подобный сценарий – основа любого прогнозирования временного ряда с той разницей, что для каждого случая приходится, как получать необходимый временной ряд посредством инструментов **Deductor Studio** (например, группировки), так и подбирать параметры очистки данных и параметры модели прогноза (например, структуры сети, если используется обучение нейронной сети, определение значимых входных факторов). В данном случае приемлемые результаты получились с настройками по умолчанию, в большинстве же случаев предстоит работа по их подбору (например, оценивая качество модели по диаграмме рассеяния).

3.6 Задание к лабораторной работе

1. Выбрать временной ряд согласно вашему варианту из файла «Временные ряды.xls».
2. Провести предобработку данных временного ряда.
3. Построить линейную авторегрессионную модель и выполнить прогнозирование.
4. Построить нейронную сеть, выполнить прогнозирование и сравнить результаты.

3.7 Контрольные вопросы

1. Какие типы прогнозов Вы можете назвать?
2. Какие этапы выполняются при решении задачи прогнозирования?
3. Дайте определение временного ряда.
4. В чем состоит задача прогнозирования временного ряда?
5. Какие «наивные» методы прогнозирования вам известны?
6. Что такое тренд временного ряда и как его получить?
7. Какие методы выделения сезонных колебаний во временных рядах вы знаете?
8. В чем особенности прогнозирования финансовых временных рядов?
9. Запишите авторегрессионную модель прогнозирования временного ряда.
10. Как выбрать глубину погружения при прогнозировании временных рядов?
11. Какую предварительную обработку данных выполняют перед построением модели прогноза?
12. Какие программные продукты можно использовать для прогнозирования? В чем их достоинства и недостатки?
13. В чем состоит преобразование данных «скользящее окно»?
14. Дайте определение нейронной сети.
15. В чем состоит задача обучения нейронной сети?
16. Какие данные необходимы для обучения нейронной сети?
17. Зачем обучают нейронную сеть?

5 ЛИТЕРАТУРА

1. Чубукова И.А. Data mining: учебное пособие – М.: Интернет-университет информационных технологий: БИНОМ: Лаборатория знаний, 2006. – 382 с. – ISBN 5-9556-0064-7.
2. Барсегян А.А., Куприянов М.С., Степаненко В.В., Холод И.И. Методы и модели анализа данных: OLAP, Data Mining. – СПб.: БХВ-Петербург, 2004. – 336с.
3. Ситник В.Ф. Інтелектуальний аналіз даних. К.: КНЕУ, 2007. –
4. А.А. Ежов, С.А. Шумский. Нейрокомпьютинг и его применения в экономике и бизнесе. –
5. Барсегян А.А., Куприянов М.С., Степаненко В.В., Холод И.И. Технологии анализа данных: Data Mining, Visual Mining, Text Mining, OLAP. – СПб.: БХВ-Петербург, 2007. – 384 с.
6. Паклин Н.Б., Орешков В.И. Бизнес-аналитика: от данных к знаниям: Учеб. пособие. 2-е изд., перераб. и доп. _ СПб.: Питер, 2010. – 704 с.