МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ
Федеральное государственное автономное образовательное учреждение высшего образования
«СЕВЕРО-КАВКАЗСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ»
Невинномысский технологический институт (филиал)

Методические указания к самостоятельным работам

студентов по дисциплине «Интеллектуальный анализ данных и машинное обучение»

(ЭЛЕКТРОННЫЙ ДОКУМЕНТ)

Направление подготовки 09.03.02 Информационные системы и технологии

Квалификация выпускника Бакалавр

Невинномысск 2022

Методические указания предназначены для студентов направления подготовки 09.03.02 Информационные системы и технологии и других технических специальностей. Они содержат рекомендации по организации самостоятельных работ студента направления для дисциплины «Интеллектуальный анализ данных и машинное обучение».

Методические указания разработаны в соответствии с требованиями ФГОС ВО в части содержания и уровня подготовки выпускников направления 09.03.02 Информационные системы и технологии.

Составитель канд. техн. наук Э.Е. Тихонов

Ответственный редактор канд. техн. наук Д.В. Болдырев

Содержание

1 Подготовка к лекциям

Главное в период подготовки к лекционным занятиям – научиться методам самостоятельного умственного труда, сознательно развивать свои творческие способности и овладевать навыками творческой работы. Для этого необходимо строго соблюдать дисциплину учебы и поведения. Четкое планирование своего рабочего времени и отдыха является необходимым условием для успешной самостоятельной работы. В основу его нужно положить рабочие программы изучаемых в семестре дисциплин.

Каждому студенту следует составлять еженедельный и семестровый планы работы, а также план на каждый рабочий день. С вечера всегда надо распределять работу на завтрашний день. В конце каждого дня целесообразно подводить итог работы: тщательно проверить, все ли выполнено по намеченному плану, не было ли каких-либо отступлений, а если были, по какой причине это произошло. Нужно осуществлять самоконтроль, который является необходимым условием успешной учебы. Если что-то осталось невыполненным, необходимо изыскать время для завершения этой части работы, не уменьшая объема недельного плана.

Слушание и запись лекций – сложный вид вузовской аудиторной работы. Внимательное слушание и конспектирование лекций предполагает интенсивную умственную деятельность студента. Краткие записи лекций, их конспектирование помогает усвоить учебный материал. Конспект является полезным тогда, когда записано самое существенное, основное и сделано это самим студентом. Не надо стремиться записать дословно всю лекцию. Такое «конспектирование» приносит больше вреда, чем пользы. Запись лекций рекомендуется вести по возможности собственными формулировками. Желательно запись осуществлять на одной странице, а следующую оставлять для проработки учебного материала самостоятельно в домашних условиях.

Конспект лекций лучше подразделять на пункты, параграфы, соблюдая красную строку. Этому в большой степени будут способствовать пункты

плана лекции, предложенные преподавателям. Принципиальные места, определения, формулы и другое следует сопровождать замечаниями «важно», «особо важно», «хорошо запомнить» и т.п. Можно делать это и с помощью разноцветных маркеров или ручек. Лучше если они будут собственными, чтобы не приходилось просить их у однокурсников и тем самым не отвлекать их во время лекции. Не лишним будет и изучение основ стенографии. Работая над конспектом лекций, всегда необходимо использовать не только учебник, но и ту литературу, которую дополнительно рекомендовал лектор. Именно такая серьезная, кропотливая работа с лекционным материалом позволит глубоко овладеть знаниями.

2 Подготовка к лабораторным и практическим занятиям

Для того чтобы лабораторные занятия приносили максимальную пользу, необходимо помнить, что упражнение и решение задач проводятся по рассмотренному на лекциях материалу и связаны, как правило, с детальным разбором отдельных вопросов лекционного курса. Следует подчеркнуть, что только после усвоения лекционного материала с определенной точки зрения (а именно с той, с которой он излагается на лекциях) он будет закрепляться студентом на лабораторных занятиях как в результате обсуждения и анализа лекционного материала, так и с помощью решения проблемных ситуаций, задач. При этих условиях студент не только хорошо усвоит материал, но и научится применять его на практике, а также получит дополнительный стимул (и это очень важно) для активной проработки лекции.

При самостоятельном решении задач нужно обосновывать каждый этап решения, исходя из теоретических положений курса. Если студент видит несколько путей решения проблемы (задачи), то нужно сравнить их и выбрать самый рациональный. Полезно до начала вычислений составить краткий план решения проблемы (задачи). Решение проблемных задач или примеров следует излагать подробно, вычисления располагать в строгом

порядке, отделяя вспомогательные вычисления от основных. Решения при необходимости нужно сопровождать комментариями, схемами, чертежами и рисунками.

Следует помнить, что решение каждой учебной задачи должно доводиться до окончательного логического ответа, которого требует условие, и по возможности с выводом. Полученный ответ следует проверить способами, вытекающими из существа данной задачи. Полезно также (если возможно) решать несколькими способами и сравнить полученные результаты. Решение задач данного типа нужно продолжать до приобретения твердых навыков в их решении.

3 Самостоятельное изучение темы. Конспект

Конспект – наиболее совершенная и наиболее сложная форма записи. Слово «конспект» происходит от латинского «conspectus», что означает «обзор, изложение». В правильно составленном конспекте обычно выделено самое основное в изучаемом тексте, сосредоточено внимание на наиболее существенном, в кратких и четких формулировках обобщены важные теоретические положения.

Конспект представляет собой относительно подробное, последовательное изложение содержания прочитанного. На первых порах целесообразно в записях ближе держаться тексту, прибегая зачастую к прямому цитированию автора. В дальнейшем, по мере выработки навыков конспектирования, записи будут носить более свободный и сжатый характер.

Конспект книги обычно ведется в тетради. В самом начале конспекта указывается фамилия автора, полное название произведения, издательство, год и место издания. При цитировании обязательная ссылка на страницу книги. Если цитата взята из собрания сочинений, то необходимо указать соответствующий том. Следует помнить, что четкая ссылка на источник – непременное правило конспектирования. Если конспектируется статья, то указывается, где и когда она была напечатана.

Конспект подразделяется на части в соответствии с заранее продуманным планом. Пункты плана записываются в тексте или на полях конспекта. Писать его рекомендуется четко и разборчиво, так как небрежная запись с течением времени становиться малопонятной для ее автора. Существует правило: конспект, составленный для себя, должен быть по возможности написан так, чтобы его легко прочитал, и кто-либо другой.

Формы конспекта могут быть разными и зависят от его целевого назначения (изучение материала в целом или под определенным углом зрения, подготовка к докладу, выступлению на занятии и т.д.), а также от характера произведения (монография, статья, документ и т.п.). Если речь идет просто об изложении содержания работы, текст конспекта может быть сплошным, с выделением особо важных положений подчеркиванием или различными значками.

В случае, когда не ограничиваются переложением содержания, а фиксируют в конспекте и свои собственные суждения по данному вопросу или дополняют конспект соответствующими материалами их других источников, следует отводить место для такого рода записей. Рекомендуется разделить страницы тетради пополам по вертикали и в левой части вести конспект произведения, а в правой свои дополнительные записи, совмещая их по содержанию.

Конспектирование в большей мере, чем другие виды записей, помогает вырабатывать навыки правильного изложения в письменной форме важные теоретических и практических вопросов, умение четко их формулировать и ясно излагать своими словами.

Таким образом, составление конспекта требует вдумчивой работы, затраты времени и труда. Зато во время конспектирования приобретаются знания, создается фонд записей.

Конспект может быть текстуальным или тематическим. В текстуальном конспекте сохраняется логика и структура изучаемого произведения, а запись ведется в соответствии с расположением материала в книге. За основу

тематического конспекта берется не план произведения, а содержание какой-либо темы или проблемы.

Текстуальный конспект желательно начинать после того, как вся книга прочитана и продумана, но это, к сожалению, не всегда возможно. В первую очередь необходимо составить план произведения письменно или мысленно, поскольку в соответствии с этим планом строится дальнейшая работа. Конспект включает в себя тезисы, которые составляют его основу. Но, в отличие от тезисов, конспект содержит краткую запись не только выводов, но и доказательств, вплоть до фактического материала. Иначе говоря, конспект – это расширенные тезисы, дополненные рассуждениями и доказательствами, мыслями и соображениями составителя записи.

Как правило, конспект включает в себя и выписки, но в него могут войти отдельные места, цитируемые дословно, а также факты, примеры, цифры, таблицы и схемы, взятые из книги. Следует помнить, что работа над конспектом только тогда будет творческой, когда она не ограничена текстом изучаемого произведения. Нужно дополнять конспект данными из других источников.

В конспекте необходимо выделять отдельные места текста в зависимости от их значимости. Можно пользоваться различными способами: подчеркиваниями, вопросительными и восклицательными знаками, репликами, краткими оценками, писать на полях своих конспектов слова: «важно», «очень важно», «верно», «характерно».

В конспект могут помещаться диаграммы, схемы, таблицы, которые придадут ему наглядность.

Составлению тематического конспекта предшествует тщательное изучение всей литературы, подобранной для раскрытия данной темы. Бывает, что какая-либо тема рассматривается в нескольких главах или в разных местах книги. А в конспекте весь материал, относящийся к теме, будет сосредоточен в одном месте. В плане конспекта рекомендуется делать пометки, к каким источникам (вплоть до страницы) придется обратиться для

раскрытия вопросов. Тематический конспект составляется обычно для того, чтобы глубже изучить определенный вопрос, подготовиться к докладу, лекции или выступлению на семинарском занятии. Такой конспект по содержанию приближается к реферату, докладу по избранной теме, особенно если включает и собственный вклад в изучение проблемы.

МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ
Федеральное государственное автономное образовательное учреждение высшего образования
«СЕВЕРО-КАВКАЗСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ»
Невинномысский технологический институт (филиал)

Методические указания для выполнения лабораторных работ по дисциплине «Интеллектуальный анализ данных и машинное обучение»

(ЭЛЕКТРОННЫЙ ДОКУМЕНТ)

Направление подготовки 09.03.02 Информационные системы и технологии

Квалификация выпускника Бакалавр

Невинномысск 2022

Методические указания предназначены для проведения лабораторных работ по дисциплине «Интеллектуальный анализ данных и машинное обучение» для студентов направления подготовки 09.03.02 Информационные системы и технологии и соответствуют требованиям ФГОС ВО направления подготовки бакалавров.

Составитель: доцент кафедры ИСЭА Э.Е. Тихонов

Содержание

ВВЕДЕНИЕ

Аналитическая платформа *DEDUCTOR* состоит из следующих пяти компонентов: *Deductor Studio, Deductor Warehouse, Deductor Viewer, Deductor Studio* и *Deductor Client*.

Deductor Warehouse – многомерное кроссплатформенное хранилище данных, аккумулирующее всю необходимую для анализа предметной области информацию. Использование единого хранилища позволяет обеспечить непротиворечивость данных, их централизованное хранение автоматически обеспечивает всю необходимую поддержку процесса анализа данных.

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.

Deductor Academic Studio предназначен только для образовательных целей. Использование данной версии в коммерческих целях запрещено. Для коммерческого применения необходимо приобрести Deductor Professional или Enterprise. Данную версию можно бесплатно получить на электронном ресурсе *http://www.basegroup.ru*

Deductor Viewer – это облегченная версия Deductor Studio, предназначенная для отображения построенных в Deductor Studio отчетов. Она не включает в себя механизмов создания сценариев, но обладает полноценными возможностями по их выполнению и визуализации результатов.

Deductor Server – сервер удаленной аналитической обработки. Он позволяет выполнять на сервере операции «прогона» данных через существующие сценарии и переобучение моделей. *Deductor Server* ориентирован на обработку больших объемов данных и работу в территориально-распределённой системе.

Лабораторная работа 1

Анализ признаков и оценка их информативности

Цель работы: ознакомиться с возможностями аналитического пакета *Deductor Academic*.

Программа работы

1. Выполнить импорт данных в программный комплекс *Deductor*.
2. Выполнить задание по предварительной парциальной обработке данных.
3. Выполнить задание по предварительной обработке путем удаления аномалий в данных.
4. Выполнить задание по предварительной обработке путем сглаживания данных методом спектральной обработки.
5. Выполнить задание по удаление шумов на этапе предварительной обработке данных.
6. Ознакомиться с возможностями автоматического анализа качества импортируемых данных.

Методические указания по выполнению работы

1.1 Импорт данных в программный комплекс Deductor Academic

Импорт данных является отправной точкой анализа данных. Импорт в *Deductor* может осуществляться из популярных форматов хранения данных, таких как *Excel*, *Access*, *MS SQL*, *Oracle*, Текстовый файл и прочих. Кроме того, имеется универсальный доступ к любому источнику данных посредством ADO или ODBC (Только в коммерческой версии, в бесплатной версии возможен импорт из *.txt, *.*csv* и *.*ded*).

Импорт данных из текстового файла с разделителями осуществляется путем вызова мастера импорта на панели «Сценарии» (рис. 1.1).

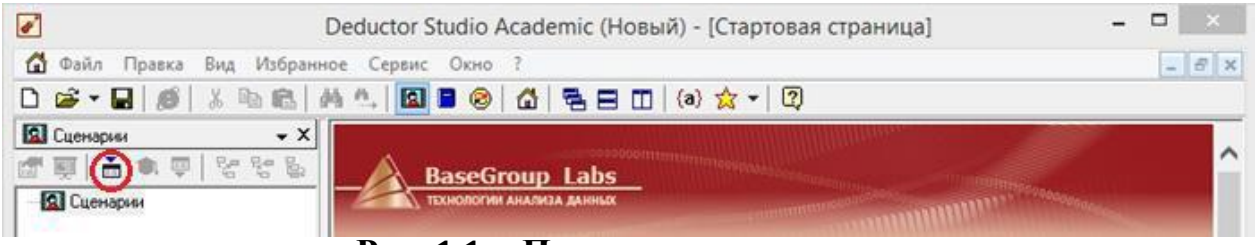


Рис. 1.1. - Панель сценарии

После запуска мастера импорта укажем тип импорта «Текстовый файл» и перейдем к настройке импорта (рис. 1.2-3). Укажем имя файла, из

которого необходимо получить данные. В окне просмотра, выбранного файла можно увидеть содержание данного файла.

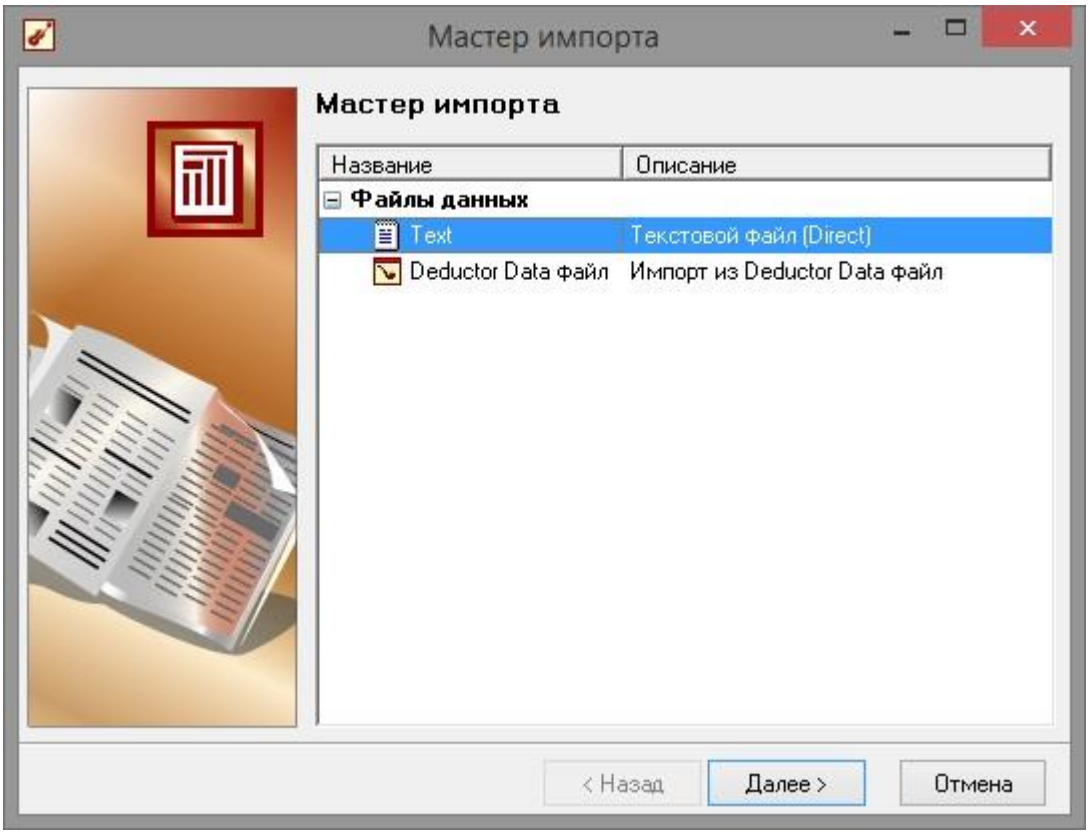


Рис. 1.2 - Мастер импорта

Далее перейдем к настройке параметров импорта (рис. 1.4). На этой странице мастера предоставляется возможность указать, с какой строки следует начать импорт, указать то, что первая строка является заголовком, возможность добавить первичный ключ. Указать, что является символом-разделителем столбцов, а также указать ограничитель строк, разделитель целой и дробной части вещественного числа, разделитель компонентов даты и ее формат.

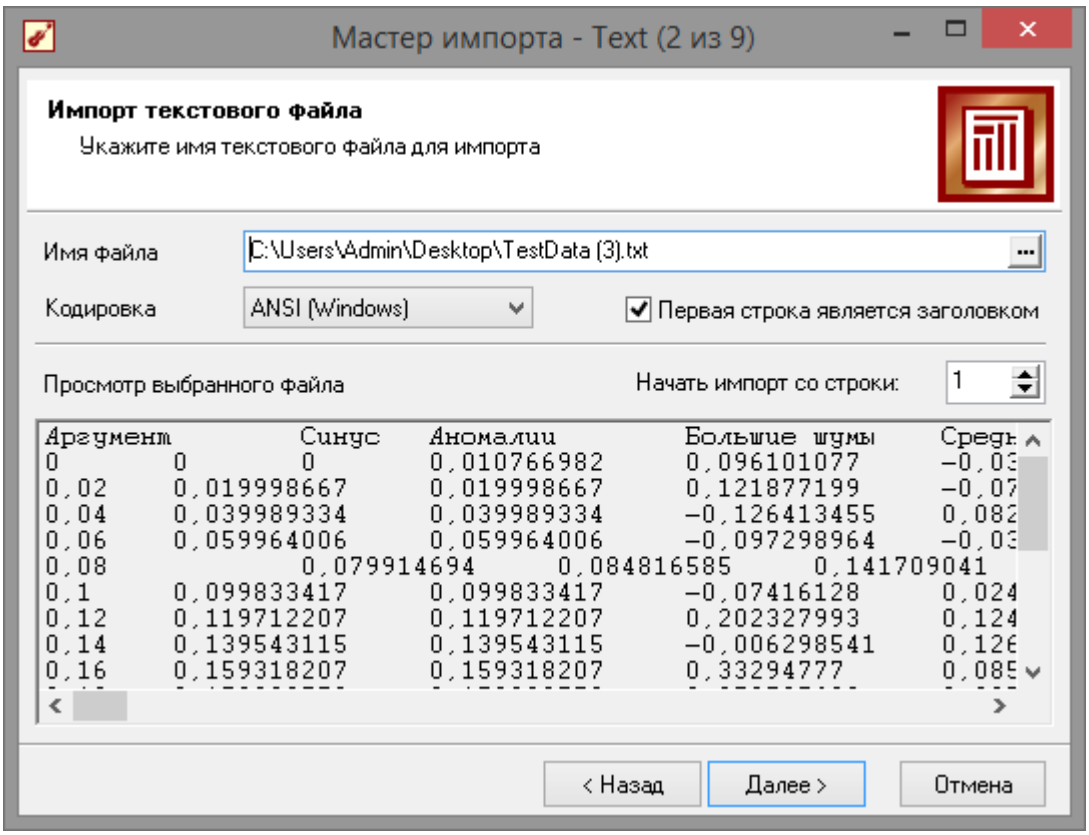


Рис. 1.3 - Выбор файла

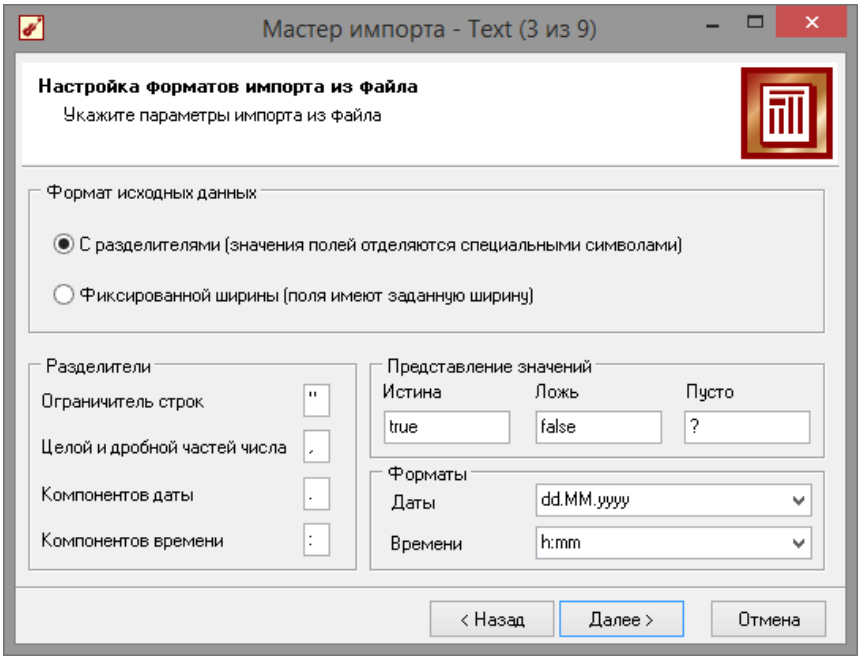


Рис. 1.4 - Параметры импорта

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. Далее перейдем к настройке свойств полей (рис. 1.5).

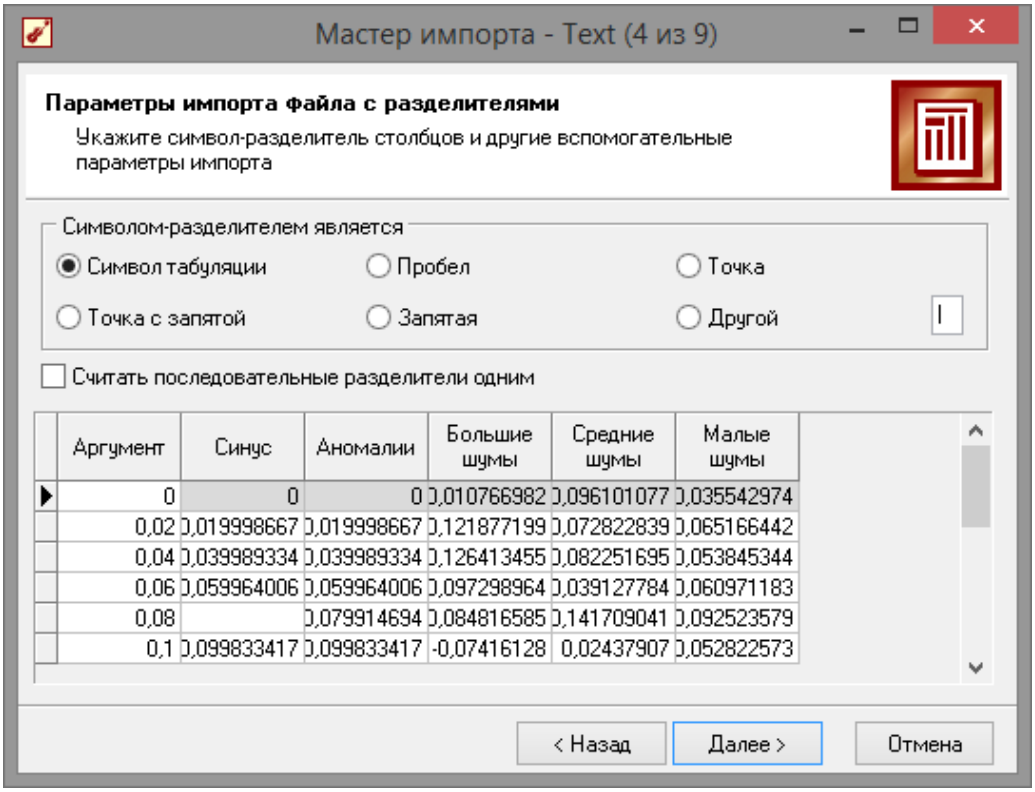


Рис. 1.5 - Параметры разделителей

На этом шаге мастера предоставляется возможность настроить имя, название (метку), размер, тип данных, вид данных и назначение (рис 1.6). Некоторые свойства (например, тип данных) можно задавать для выделенного набора столбцов. Вид данных определяет – конечный ли это набор (дискретные) или бесконечный (непрерывные). Назначение столбцов определяет характер их использования в алгоритмах обработки (при импорте можно оставить значение по умолчанию). Необходимо убедиться, что в данном случае тип данных у все столбцов выставлен как вещественный, так как в более старых версиях тип определялся по первой строке, а в демо-примере столбцы «Аргумент», «Синус» и «Аномалии» имеют в первой строке значение «0», что могло приводить к неправильному определению типа данных.

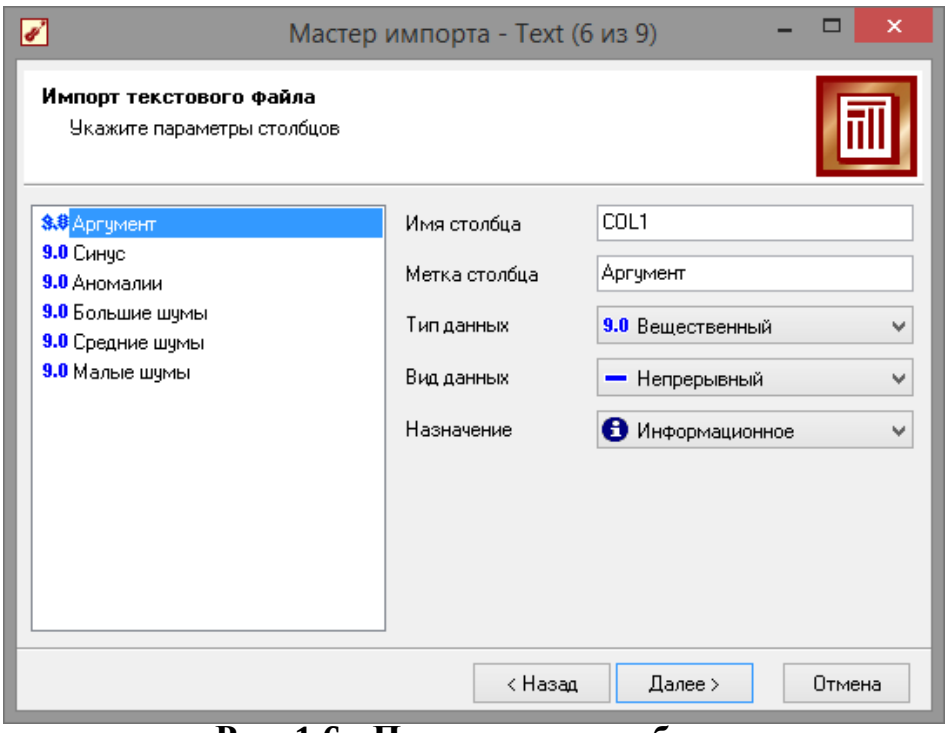


Рис. 1.6 - Параметры столбцов

Далее осталось только выполнить импорт данных, нажав на кнопку «Пуск» на следующем шаге мастера импорта (рис. 1.7). После импорта данных на следующем шаге мастера необходимо выбрать способ отображения данных (рис. 1.8). В данном случае самым информативным является диаграмма, выберем ее.

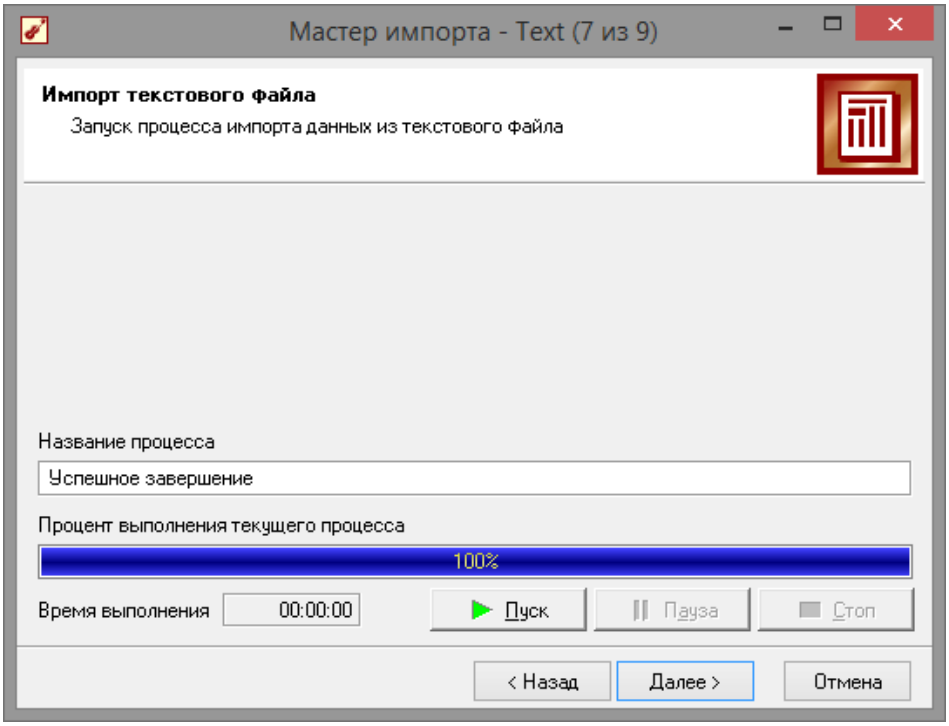


Рис. 1.7 - Импорт файла

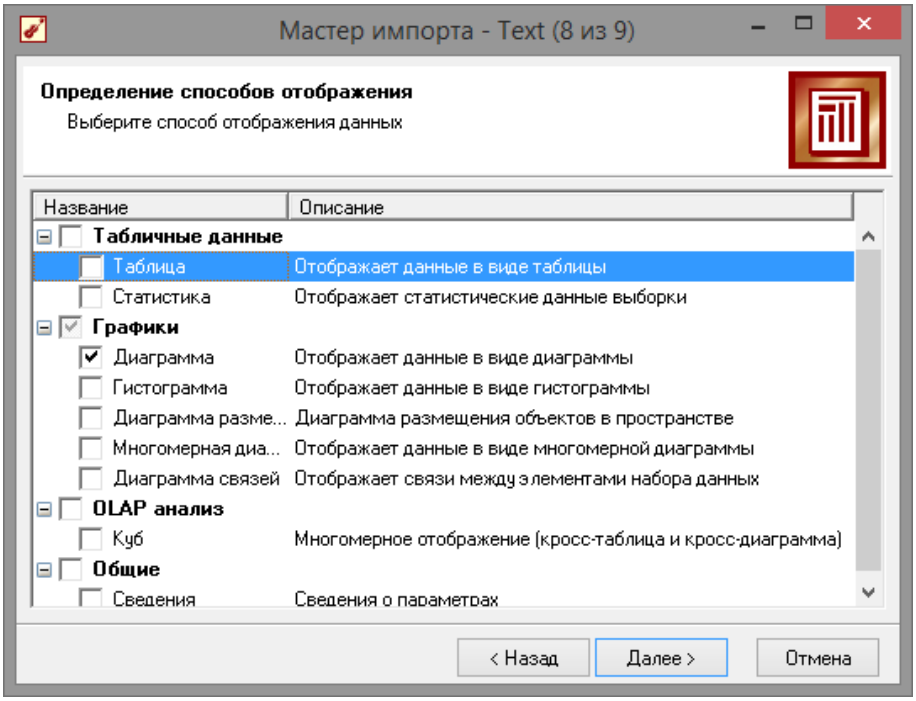


Рис. 1.8 - Способ отображения

От того, какие способы отображения будут выбраны на этом этапе, зависят последующие шаги мастера. В данном случае необходимо настроить, какие столбцы диаграммы следует отображать и как именно. Выберем для отображения поле «СИНУС» и тип диаграммы «Линии» (рис. 1.9).

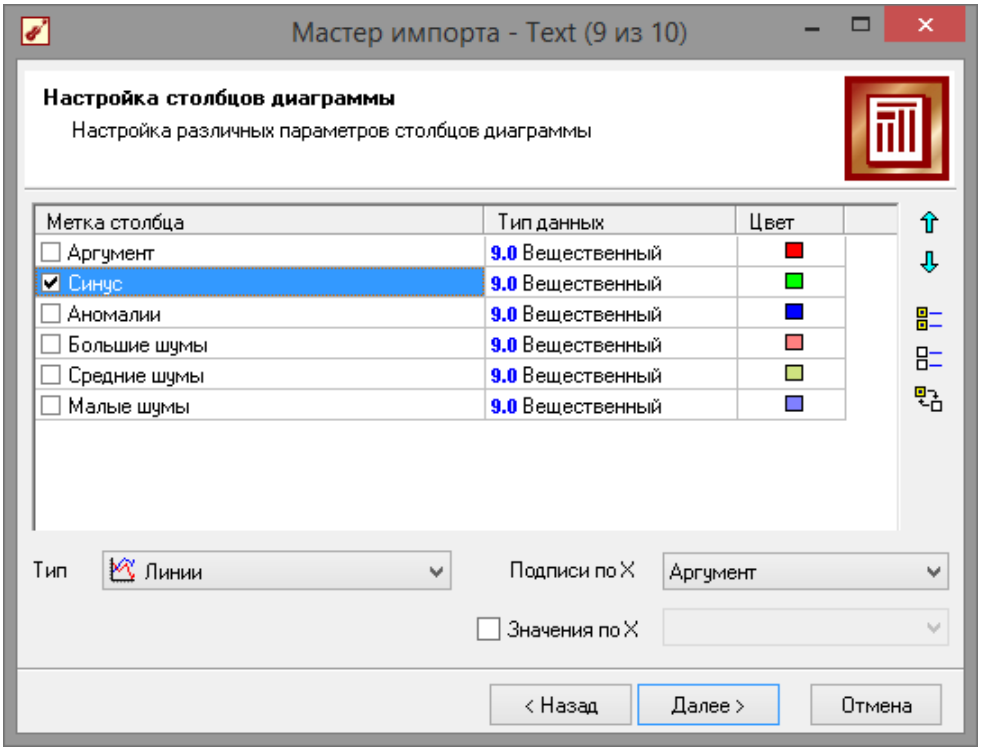


Рис. 1.9 - Настройка столбцов

На последнем шаге мастера необходимо указать название ветки в дереве сценариев. Напишем в поле заголовка окна «Импорт примера для демонстрации предобработки данных» и нажмем «Готово» (рис. 1.10). На этом работа мастера импорта заканчивается. Теперь в дереве сценариев появится новый узел с необходимыми данными. В главном окне программы представлены все выбранные отображения данных этого узла. В данном случае только диаграмма. Примечание: для отображения диаграммы в 3D-виде, необходимо нажать кнопку

«3-х мерный вид» в левом верхнем углу панели «Диаграмма». А для просмотра другой диаграммы, нажать на значок лупы «Отображать поля» (рис. 1.11).

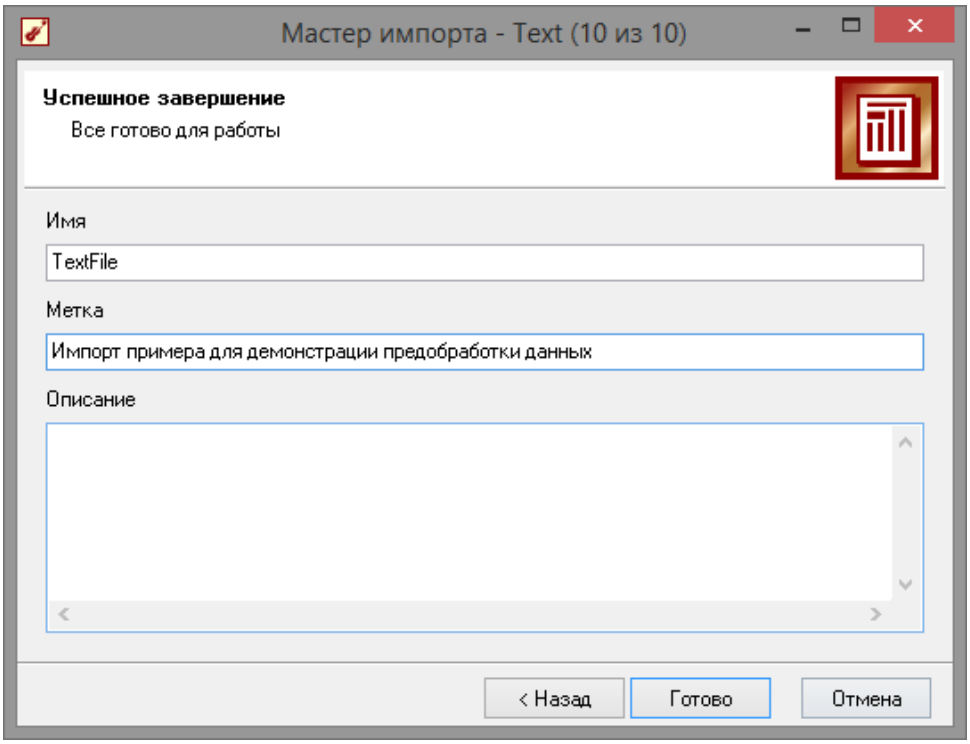


Рис. 1.10 – Завершение импорта

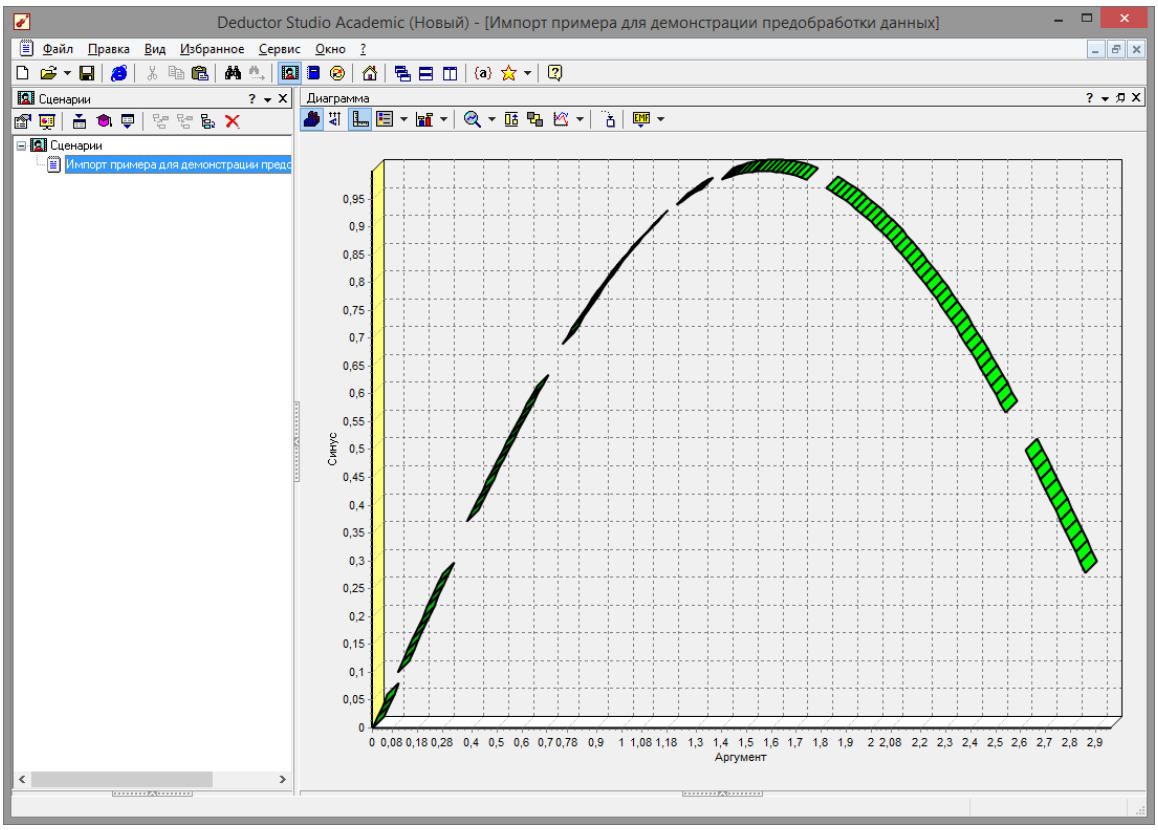


Рис. 1.11 - Диаграмма синуса

1.2 Предварительная парциальная обработка

Часто исходные данные для анализа не годятся, а качество данных влияет на качество результатов, поэтому вопрос подготовки данных для последующего анализа является очень важным. Обычно

«сырые» данные содержат в себе различные шумы, за которыми трудно увидеть общую картину, а также аномалии – влияние случайно, либо редко происходивших событий. Очевидно, что влияние этих факторов на общую модель необходимо минимизировать, т.к. модель, учитывающая их, получится неадекватной.

Парциальная предобработка служит для восстановления пропущенных данных, редактирования аномальных значений и спектральной обработке данных (например, сглаживания данных). Именно этот шаг часто проводится в первую очередь.

Рассмотрим применение обработки на примере данных из файла

«TestData.txt». Он содержит таблицу со следующими полями:

«АРГУМЕНТ» – аргумент, «СИНУС» – значения синуса аргумента (некоторые значения пустые), «АНОМАЛИИ» – синус с выбросами,

«БОЛЬШИЕ ШУМЫ» – значения синуса с большими шумами,

«СРЕДНИЕ ШУМЫ» – значения синуса со средними шумами,

«МАЛЫЕ ШУМЫ» – значения синуса с малыми шумами. Все данные можно увидеть на диаграмме после импорта из текстового файла.

Часто бывает так, что в столбце некоторые данные отсутствуют в силу каких-либо причин (данные не известны, либо их забыли внести и т.п.). Обычно из–за этого пришлось бы убрать из обработки все строки, которые содержат пропущенные данные. Но механизмы Deductor Studio позволяют решить эту проблему. Один из шагов парциальной обработки как раз отвечает за восстановление пропущенных значений. Если данные упорядочены (например, по времени), то рекомендуется в качестве восстановления пропущенных значений использовать аппроксимацию. Алгоритм сам подберет значение, которое должно стоять на месте пропущенного значения, основываясь на близлежащих данных. Если же данные не упорядочены, то следует использовать режим максимального правдоподобия, когда алгоритм подставляет вместо пропущенных данных наиболее вероятные значения, основываясь на всей выборке.

Для демонстрации воспользуемся мастером заполнения пропусков. Импортировав файл можно увидеть, что в столбце

«СИНУС» содержатся пустые значения. На диаграмме выше видно, что некоторые значения синуса пропущены. Для дальнейшей обработки необходимо из восстановить. Для этого следует запустить мастер заполнения пропусков. На рис.1.12 показаны различные варианты столбцов: с пропущенными данными; с аномалиями (выбросами); с большими шумами; со средними шумами; с малыми шумами.

Для запуска мастера необходимо выделит нужный сценарий и нажать F7, либо правый клик по необходимому сценарию откроет контекстное меню, где так же можно выбрать мастер обработки. Поскольку данные в исходном наборе упорядочены, на следующем шаге мастера обработки поставим галочку – «обрабатывать как упорядоченный набор» (рис. 1.13).

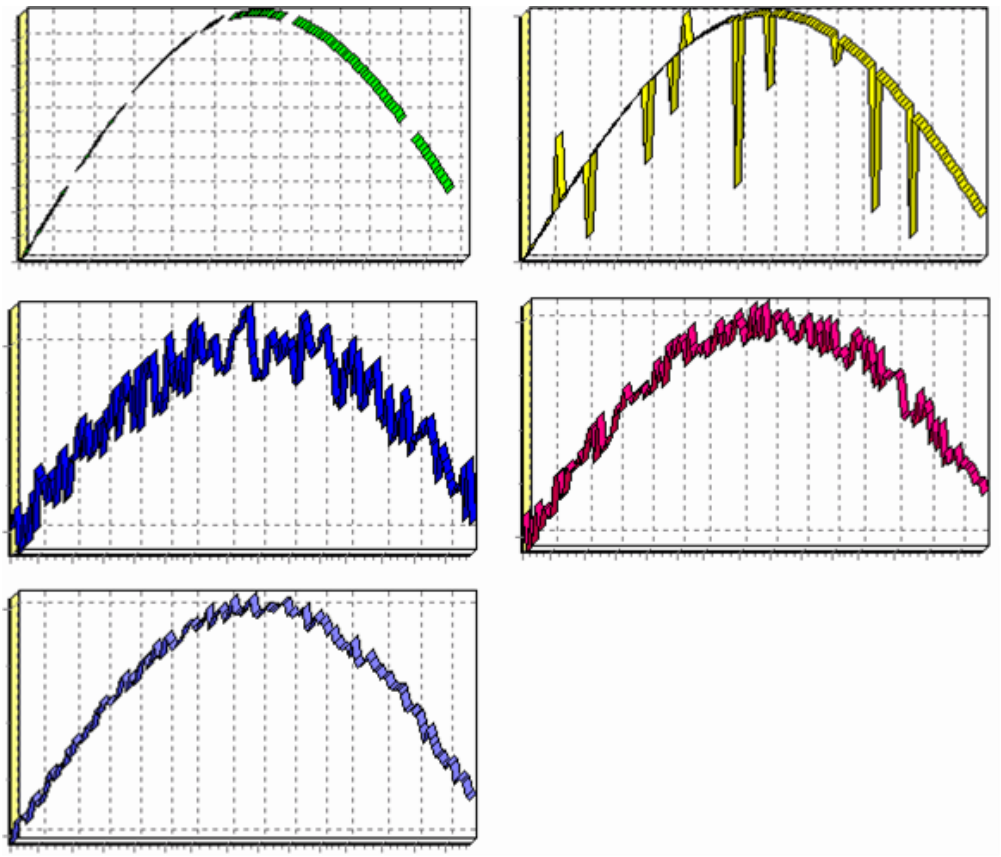


Рисунок 1.12 - Варианты столбцов

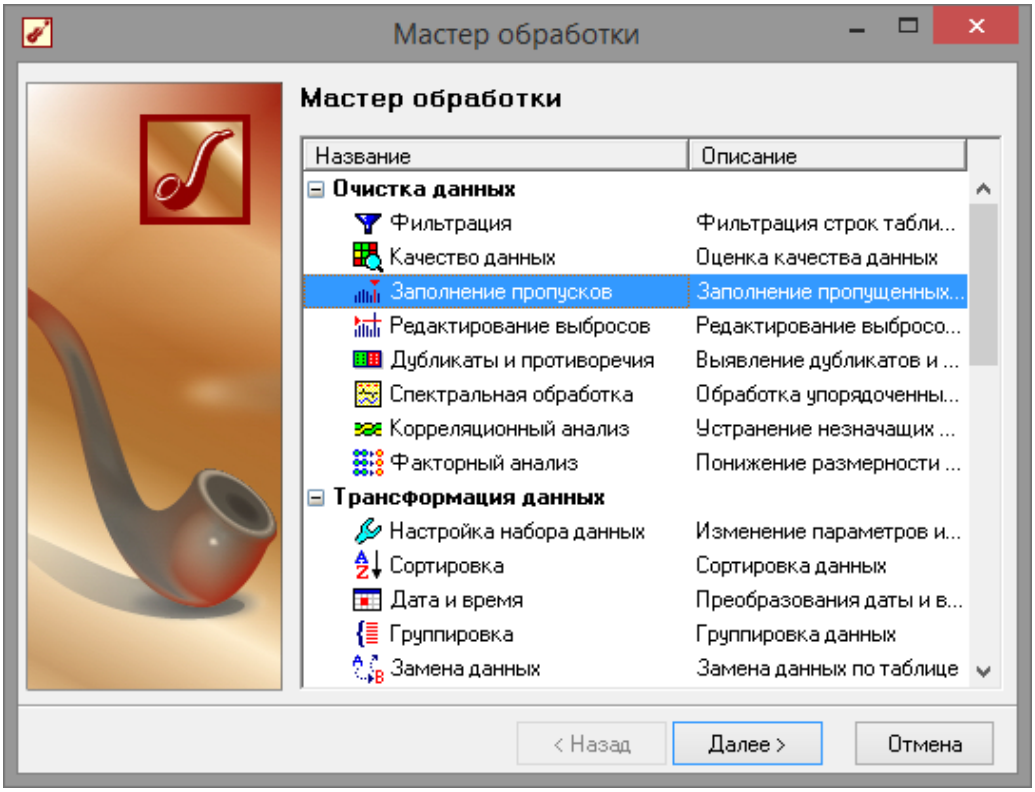


Рисунок 1.13 - Мастер обработки

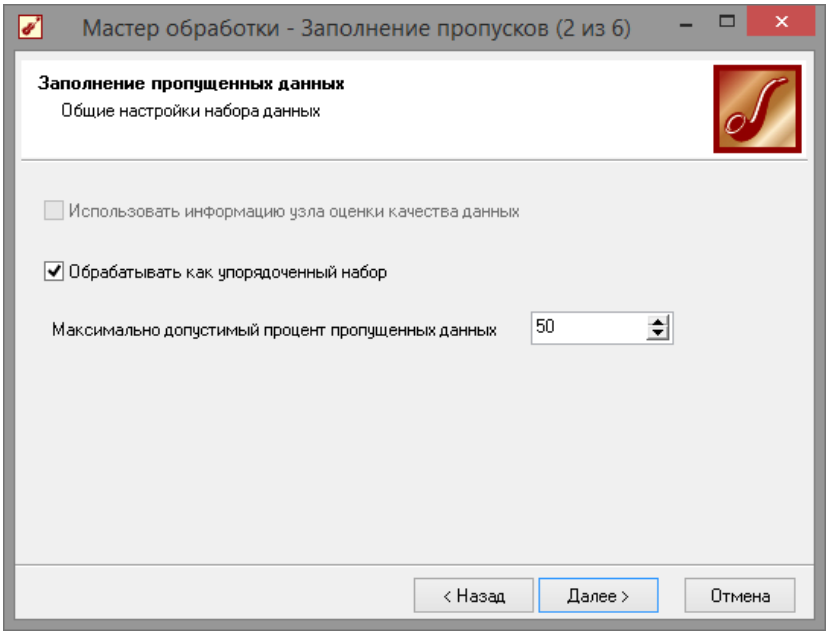


Рис. 1.14 - Мастер заполнения пропусков

Далее следует выбрать необходимый столбец и метод заполнения, в данном случае интерполяция (рис. 1.14). Перейдя на страницу запуска процесса обработки, выполняем ее, нажав на пуск, и далее выбираем тип визуализации обработанных данных (как в примере импорта) (рис. 1.15). После выполнения процесса обработки на диаграмме видно, что пропуски в данных исчезли, что и было необходимо сделать (рис. 1.16).

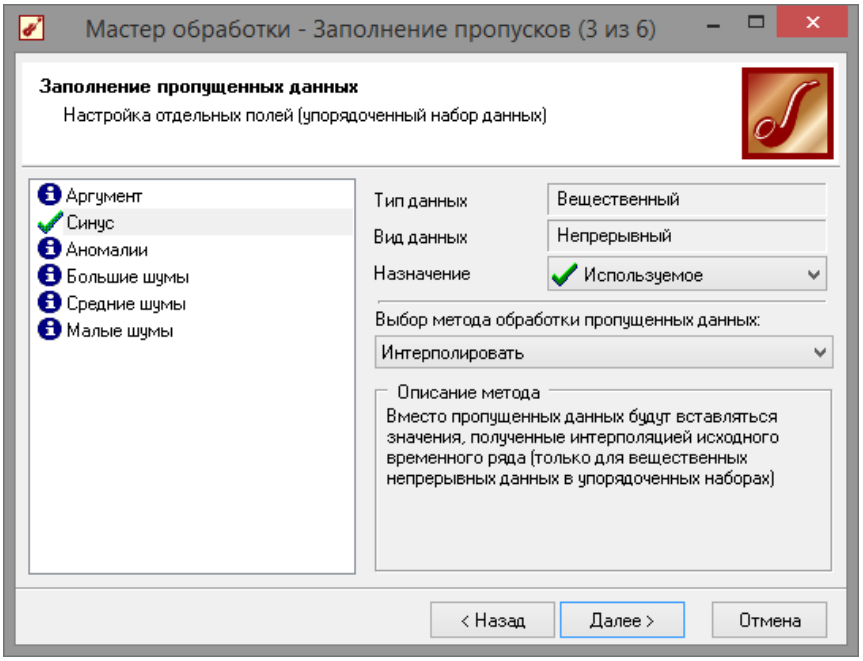


Рис. 1.15 - Настройки мастера заполнения пропусков

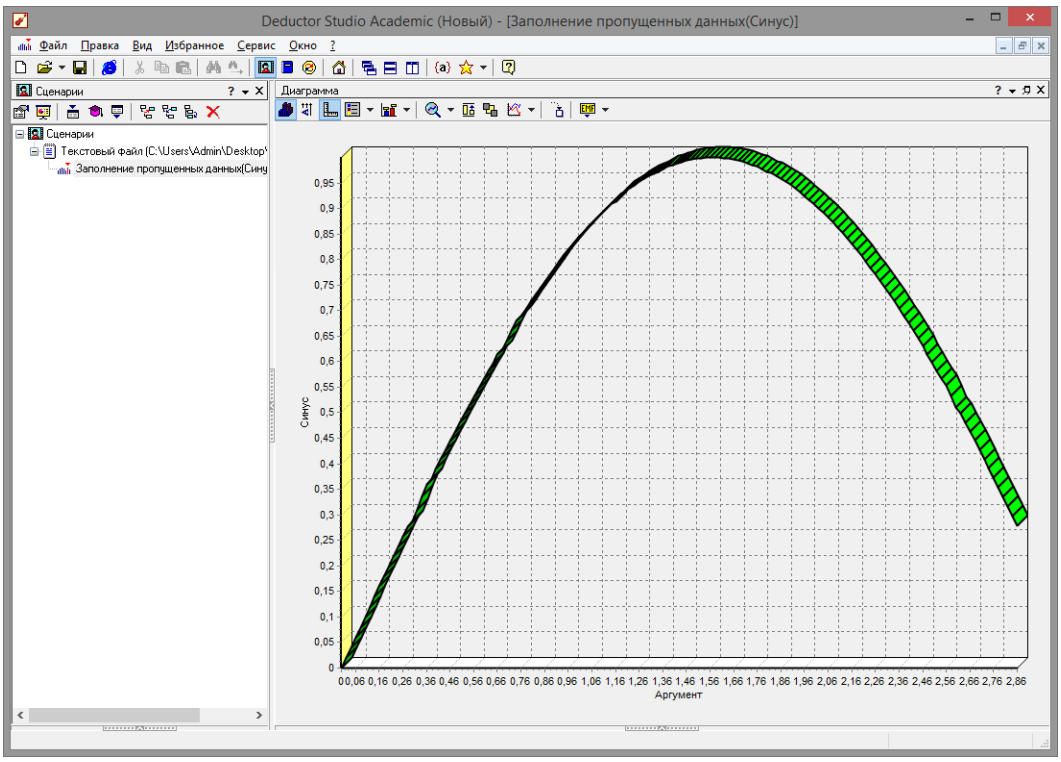


Рис.1.16 - Заполнение пропусков методом интерполяции

1.3 Удаление аномалий на этапе предобработки данных

Аномалии встречаются в «сырых» данных не реже шумов. По существу, они вообще не должны оказывать никакого влияния на результат. Если же они присутствуют при построении модели, то оказывают на нее весьма большое влияние и их предварительно необходимо устранить. Также они портят статистическую картину распределения данных. К примеру, вот как выглядят данные с аномалиями, а также гистограмма их распределения (рис. 1.17).

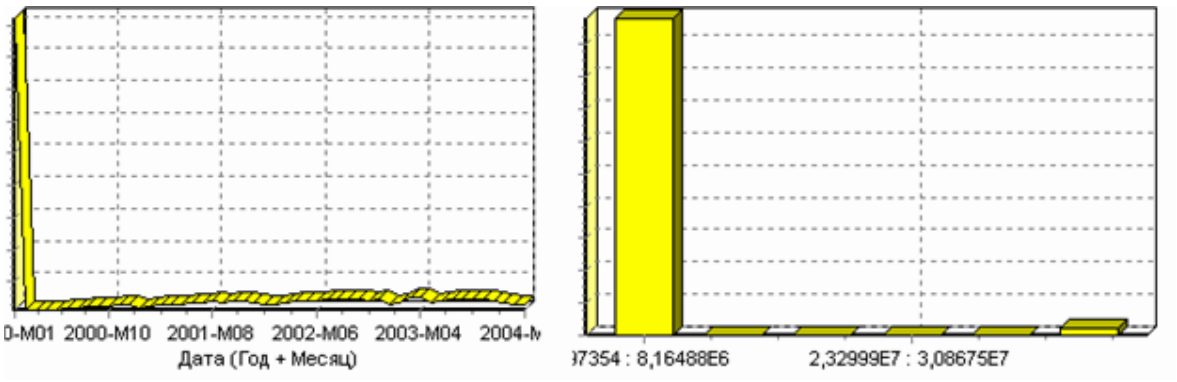


Рис. 1.17 -Гистограмма с аномалиями

Очевидно, что аномалии не позволяют определить, как характер самих данных, так и статистическую картину. После устранения аномалий те же данные представляются как показано на рис. 1.18

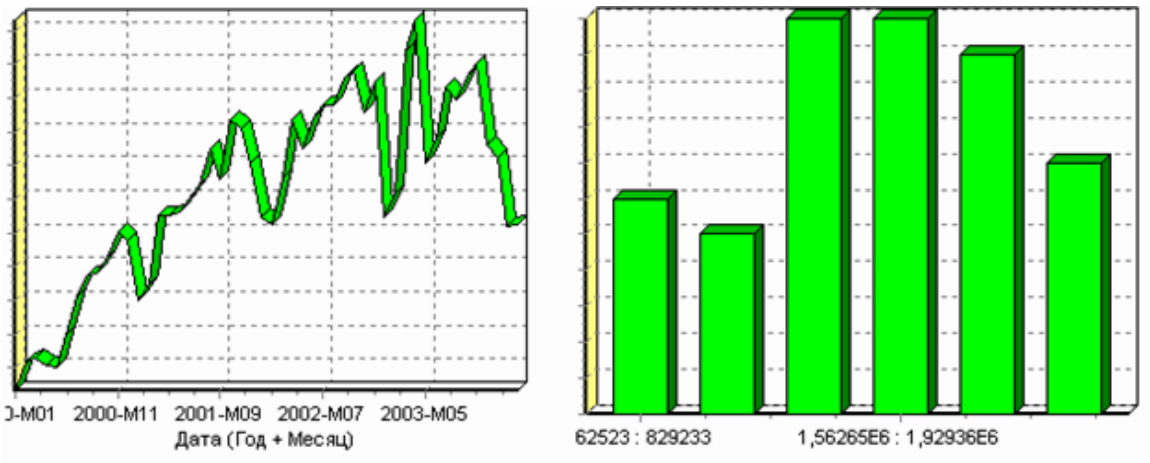


Рис. 1.18 - Гистограмма без аномалий

Следует открыть мастер обработки и выбрать редактирование выбросов (рис. 1.19). Поставить галочку - «Обрабатывать как упорядоченный набор данных» (рис. 1.20).

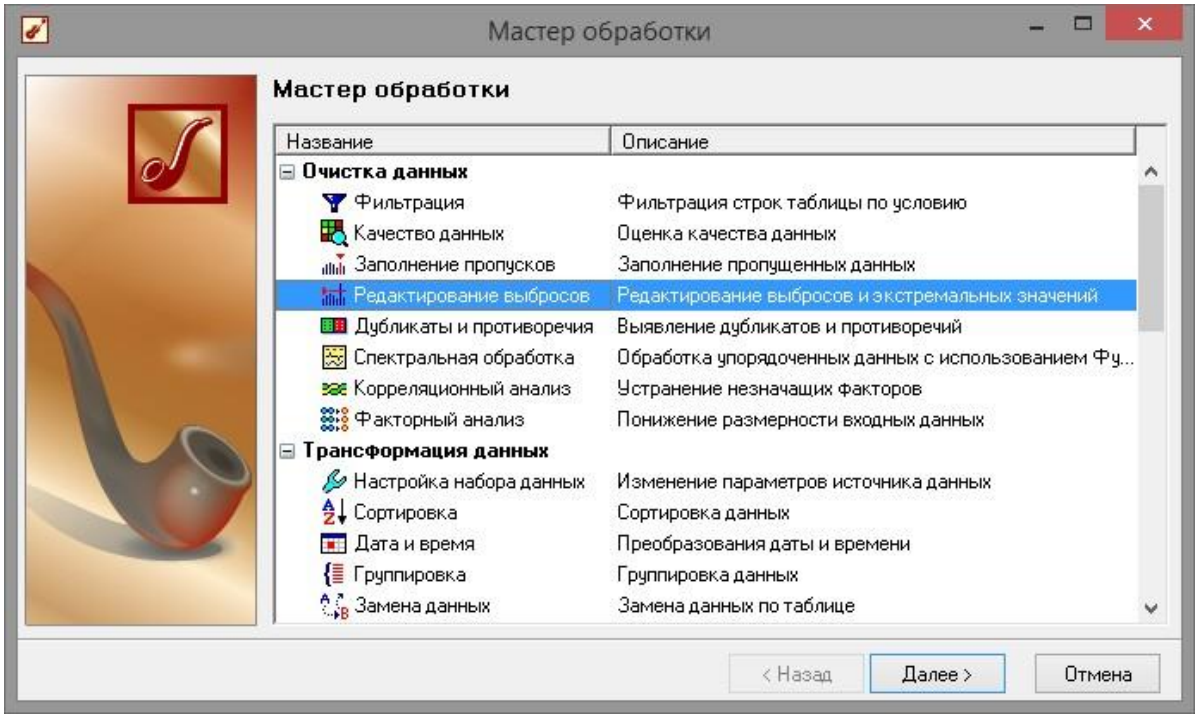


Рис. 1.19 - Мастер редактирования выбросов

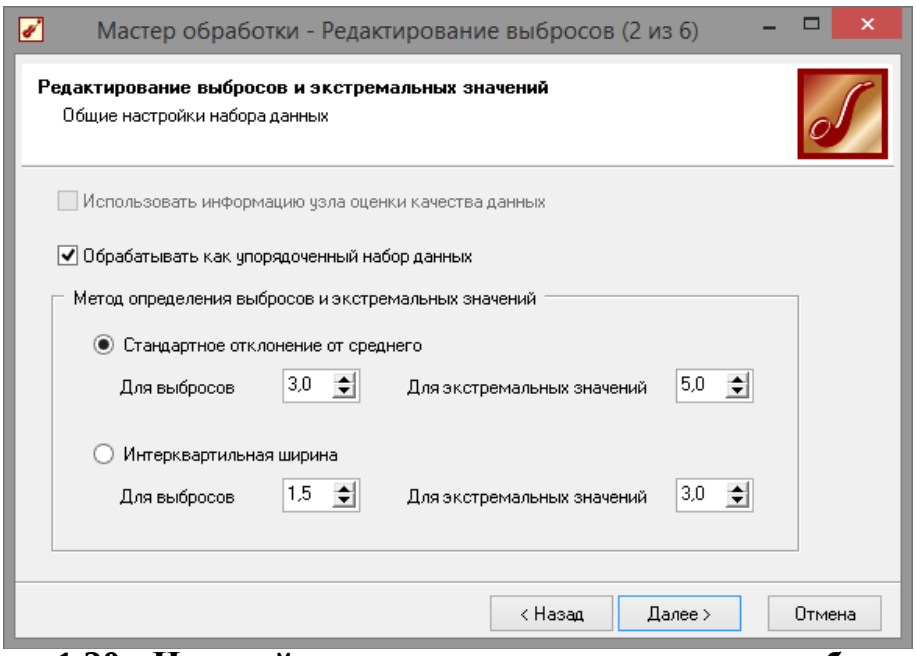


Рис. 1.20 - Настройки мастера редактирования выбросов

На следующем шаге (рис. 1.21) необходимо выбрать назначение «Используемое» только для необходимого столбца данных. И выставить большую степень подавления, так как выбросы существенны. Для множественного выделения столбцов можно использовать клавишу *Shift* и *Ctrl*.

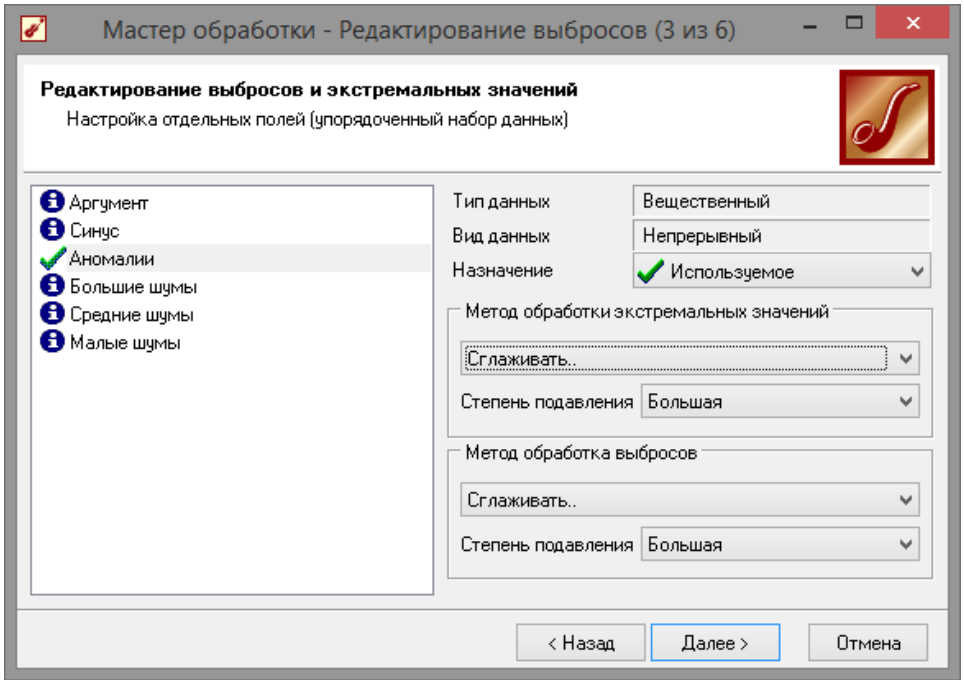


Рис. 1.21 - Настройки отдельных полей мастера редактирования выбросов

Далее нажать кнопку «Пуску» и выбрать данные для отображения как в предыдущих пунктах. После выполнения процесса обработки на диаграмме

видно, что выбросы исчезли, остались лишь небольшие возмущения, которые легко сгладить при помощи спектральной обработки (рис. 1.22).

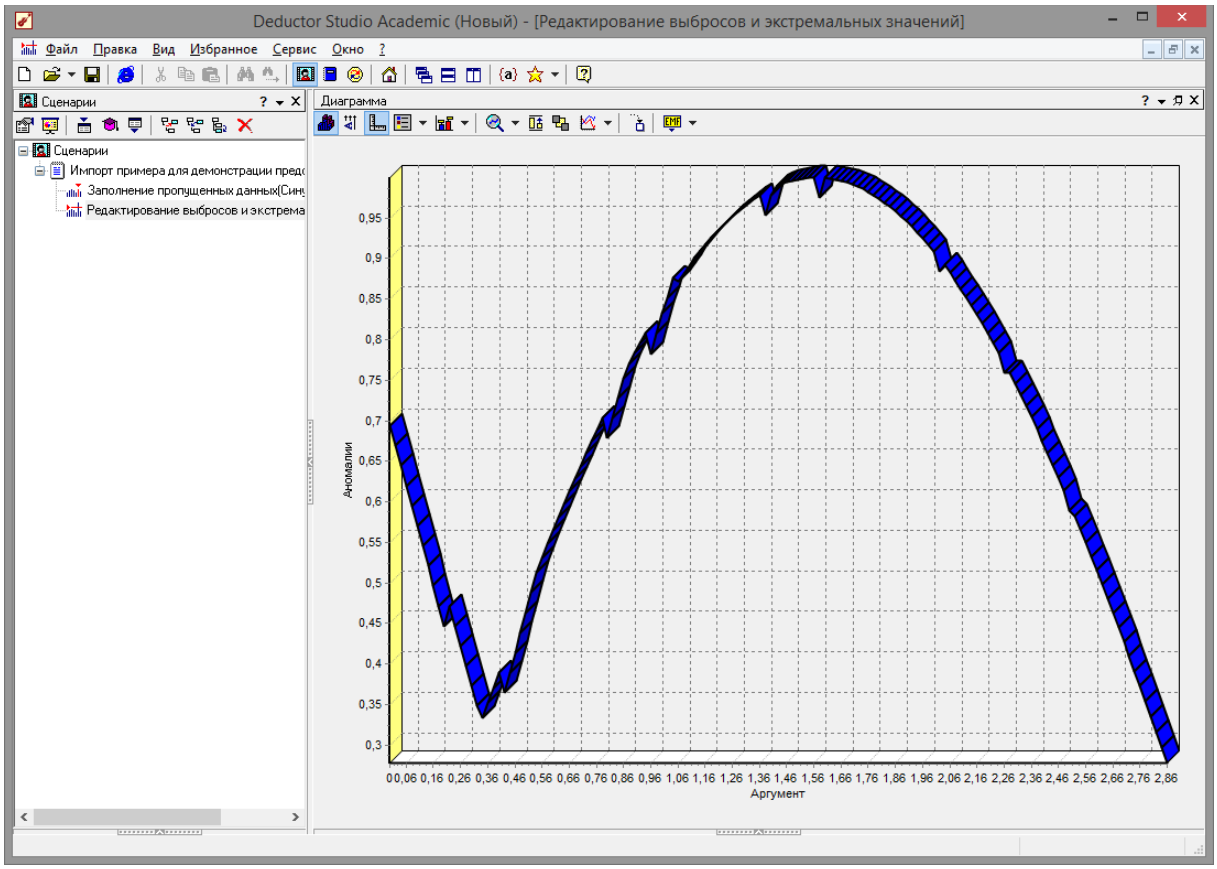


Рис. 1.22 - Диаграмма после удаления аномалий

1.4 Сглаживания данных методом спектральной обработки

Сглаживание данных применяется для удаления шумов из исходного набора, а также для выделения тенденции, которая трудно видна 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 вейвлет преобразование путем указания глубины разложения и порядка вейвлета.

Для выбора способа вейвлет преобразования открыть мастер обработки для сценария «Редактирование выбросов и экстремальных значений». В мастере следует выбрать пункт «спектральная обработка» (рис. 1.23)

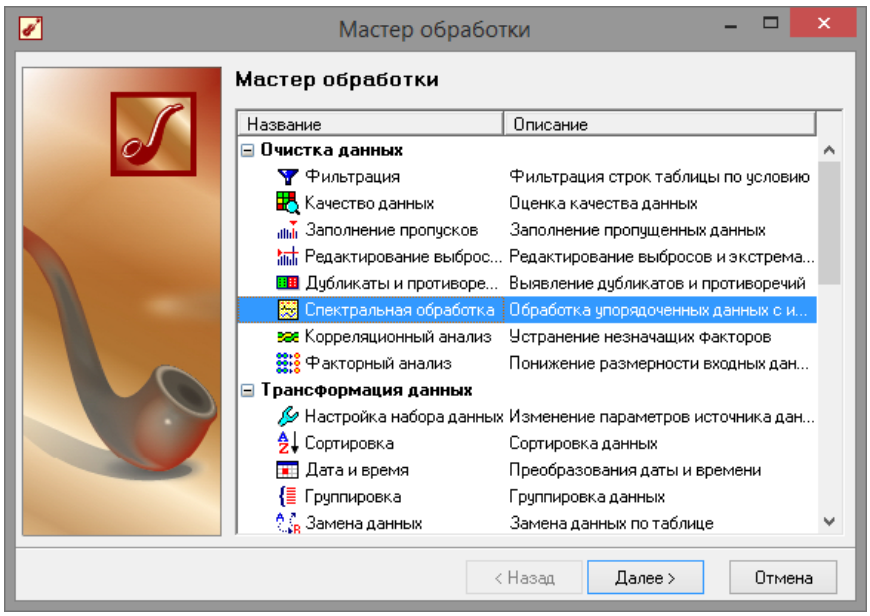


Рис. 1.23 - Мастер обработки

В мастере спектральной обработки необходимо выбрать назначение, используемое для столбца «Аномалии», и в качестве метода сглаживания данных выбрать вейвлет преобразование (рис. 1.24).

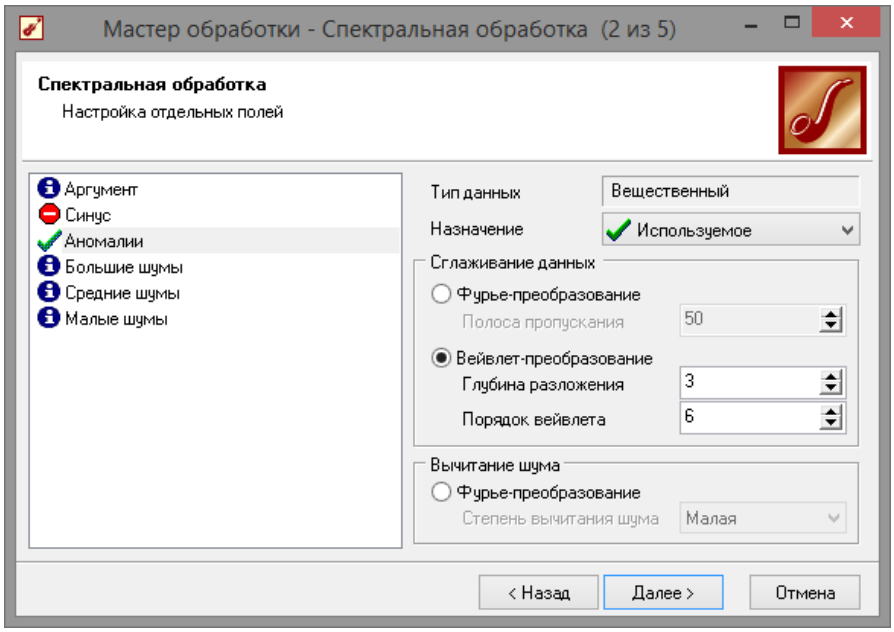


Рис. 1.24 - Мастер спектральной обработки

Далее выбрать визуализацию «Диаграмма» и столбец «Аномалии». После обработки можно убедиться на диаграмме в отсутствии выбросов (рис. 1.25).

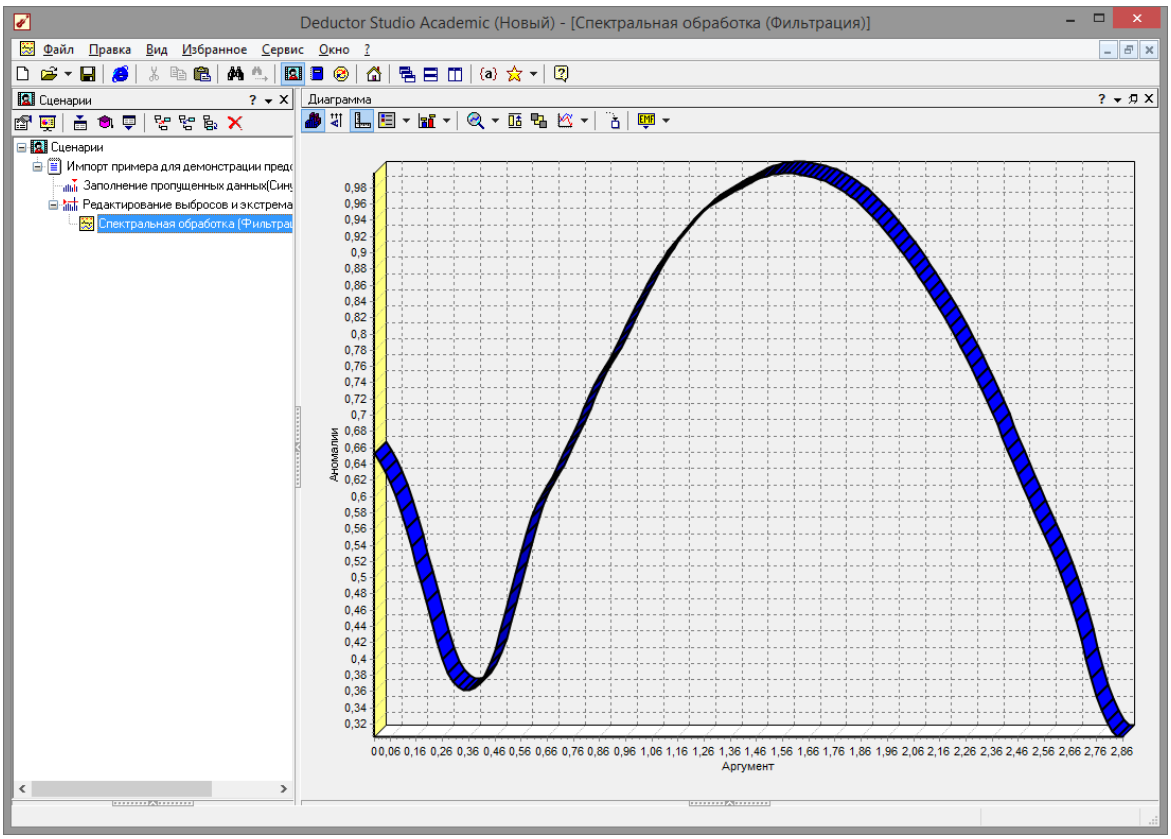


Рис. 1.25 - Диаграмма после применения спектральной обработки

1.5 Удаление шумов на этапе предварительной обработке данных

Шумы в данных не только скрывают общую тенденцию, но и проявляют себя при построении модели прогноза. Из-за них модель может получиться с плохими обобщающими качествами.

В примере по парциальной обработке есть 3 столбца с шумами:

«БОЛЬШИЕ ШУМЫ», «СРЕДНИЕ ШУМЫ», и «МАЛЫЕ ШУМЫ» -

соответственно синус с большими, средними и малыми шумами. Ясно, что для дальнейшей работы с данными эти шумы необходимо устранить. Спектральная обработка позволяет сделать это с помощью указания для этих полей в качестве типа обработки «Вычитание шума». Настройки обладают определенной гибкостью. Так, существует большая, средняя и малая степень вычитания шума. Аналитик может подобрать степень, устраивающую его.

В мастере спектральной обработки (рис. 1.26) по очереди выбрать поля «БОЛЬШИЕ ШУМЫ», «СРЕДНИЕ ШУМЫ» и

«МАЛЫЕ ШУМЫ», задать тип обработки «Вычитание шума» и

указать степень подавления – «большая», «средняя» и «малая» соответственно. В некоторых случаях неплохие результаты удаления шумов дает вейвлет преобразование. Повысить качество сглаживания шумов таким способом можно, путем подбора удовлетворительных параметров обработки (рис. 1.27).

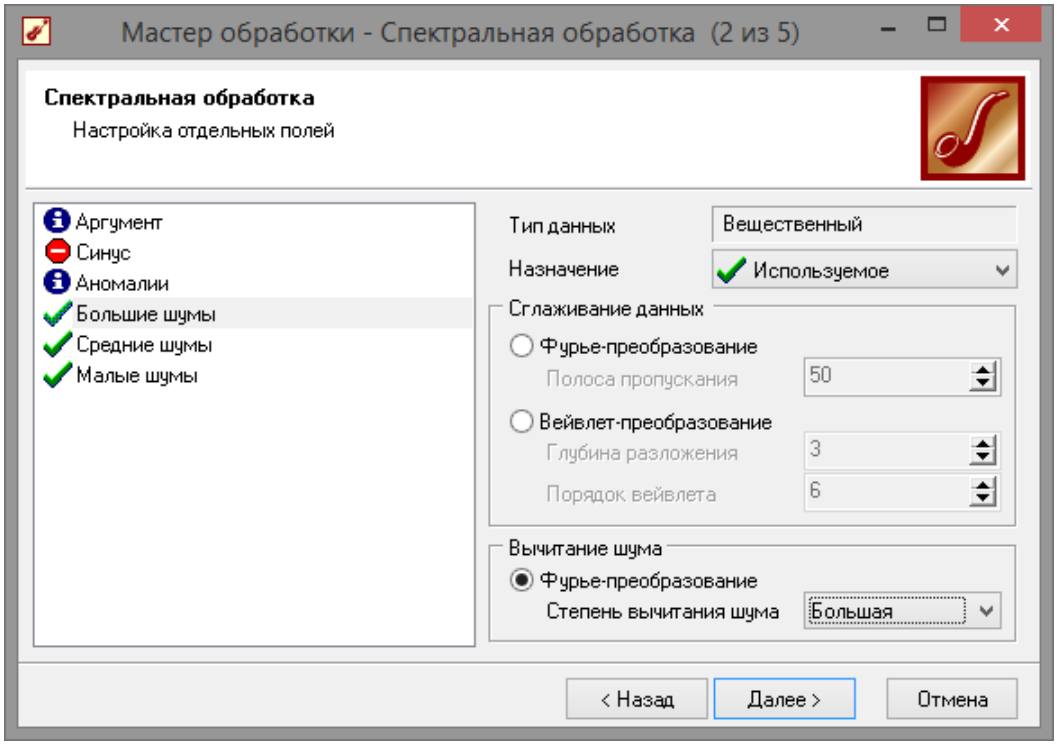


Рис. 1.26 - Настройки мастера спектральной обработки

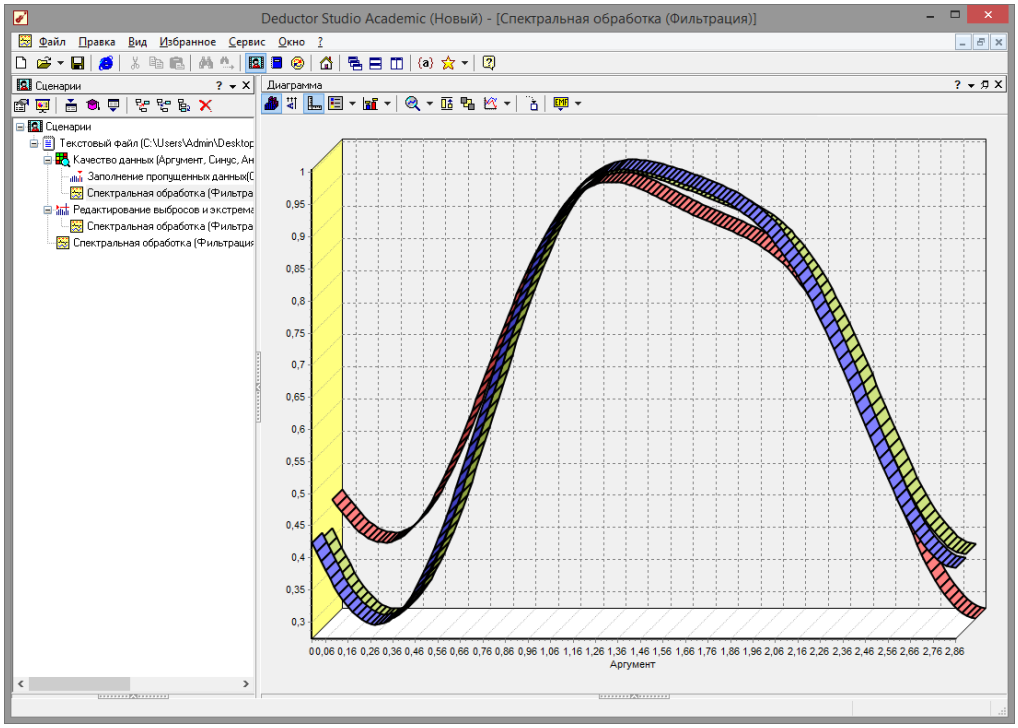


Рис 1.27 - Диаграмма после применения спектральной обработки

1.6 Возможности автоматического анализа качества импортируемых данных в *Deductor Academic*

В мастере обработки выбрать пункт «качество данных» (рис. 1.28).

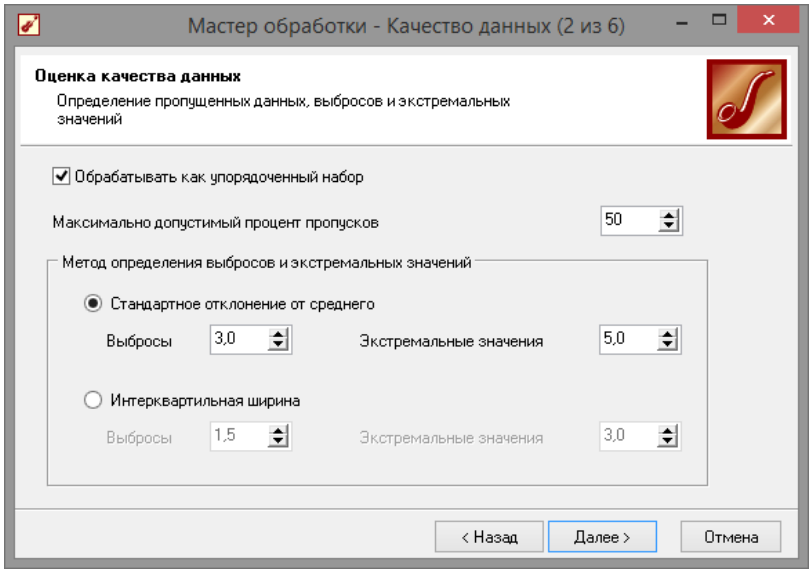


Рис. 1.28 - Мастер качества данных

После анализа мастер дает рекомендации к обработке данных и возможность автоматического исправления (рис. 1.29). Следует отметить, что автоматические исправление далеко не всегда дает желаемые результаты (рис. 1.30).

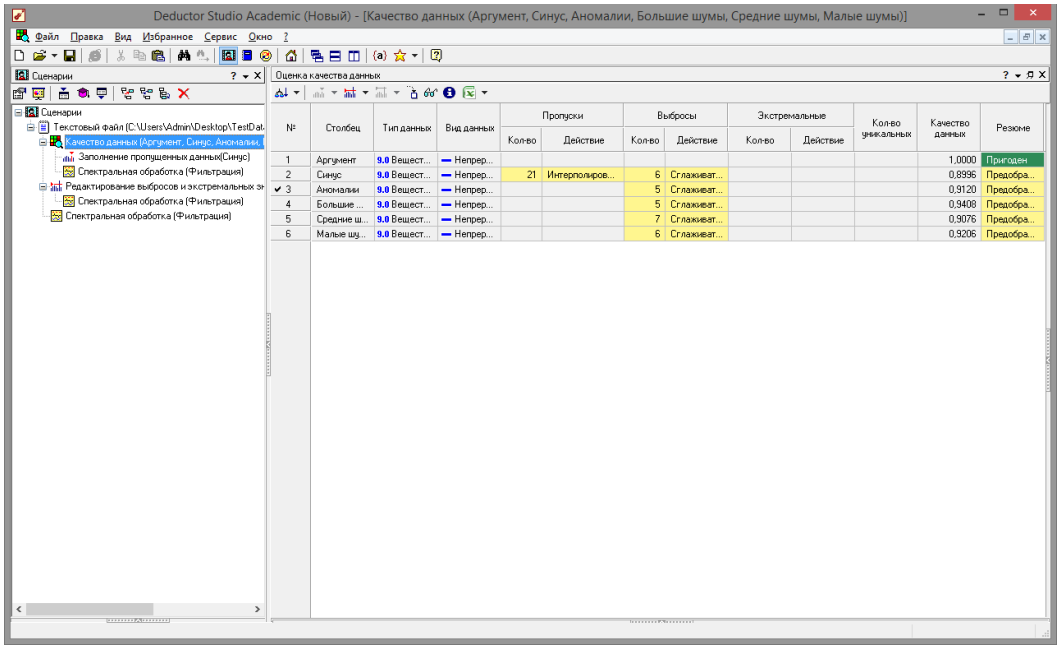


Рис. 1.29 - Результаты мастера качества до обработки данных

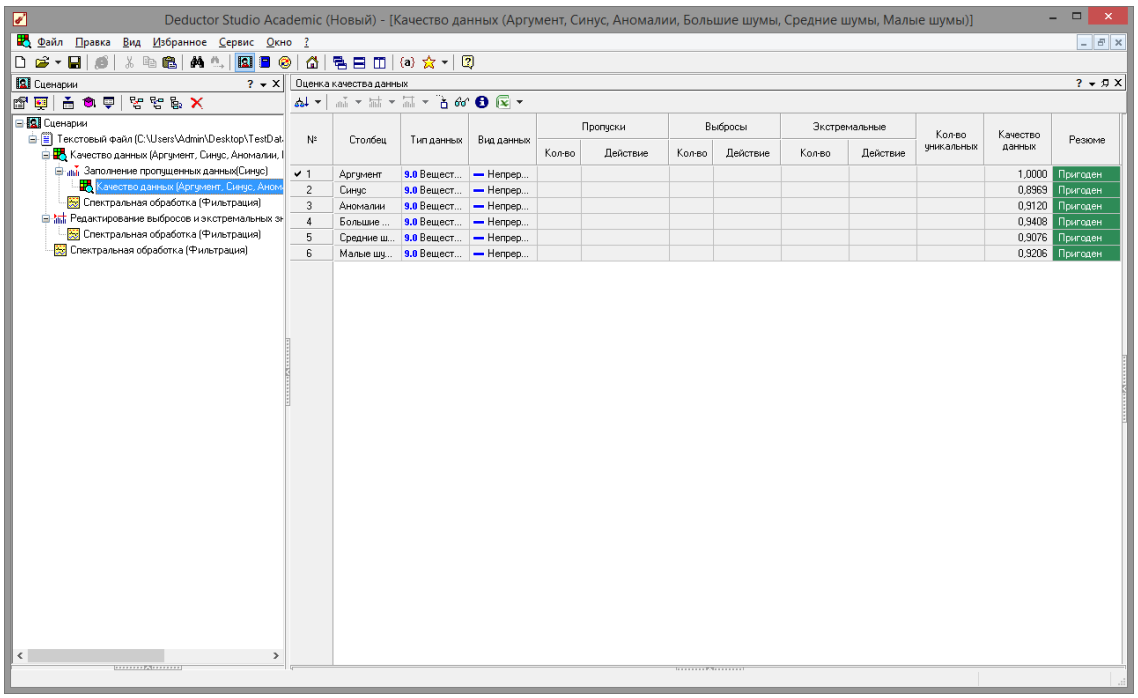


Рис. 1.30 - Вывод мастера после всех внесенных нами изменений

«Грязные данные» представляют собой очень большую проблему. Фактически они могут свести на нет все усилия по анализу данных. Причем, речь идет не о разовой операции, а о постоянной работе в этом направлении. Чисто не там, где не сорят, а там, где убирают.

Описанные выше варианты решения проблем не единственные. Есть еще достаточно много методов обработки, начиная от экспертных систем и заканчивая нейросетями. Главное, суметь грамотно ими воспользоваться. Обязательно нужно учитывать то, что методы очистки сильно привязаны к предметной области. От сферы деятельности организации и назначения хранилища данных зависит практически все. То, что для одних является шумом для других очень ценная информация. Если у нас будет априорная информация о задаче, то качество очистки данных можно увеличить на порядки.

1.7 Задание на самостоятельную работу

Сгенерировать собственный набор данных провести его анализ и выполнить предобработку данных. Данные сгенерировать в электронной таблице с помощью *MS Excel* с использованием формул и автозаполнений. Готовый файл необходимо сохранить в формате «*.*csv*» (*MS-DOS*) и при импорте в *Deductor*, выбрать в качестве разделителя «Точка с запятой».

Содержание отчёта

1. Описание предметной области

2. Зашумлённые данные (90-100 векторов размерностью не менее 4)
3. Скорректированные данные
4. Краткий порядок обработки
5. Выводы

Контрольные вопросы

1. Для чего служит программа *Deductor Academic*?
2. Зачем нужна предобработка данных?
3. Что такое парциальная предобработка?
4. Что такое вейвлет?
5. Какие данные можно импортировать в программу?

Лабораторная работа 2

Базовые методы интеллектуального анализа данных

Продолжительность работы – 8 час.

Цель работы: ознакомиться с возможностями классификации данных с помощью аналитического пакета *Deductor Academic*.

Программа работы

1. Выполнить классификацию данных с использованием алгоритма *g-mean*.
2. Выполнить классификацию данных с использованием алгоритма *k-mean*.
3. Выполнить классификацию данных с использованием нейронной сети.

Методические указания по выполнению работы

Задача разбиения множества объектов или наблюдений на априорно заданные группы, называемые классами, внутри каждой из которых они предполагаются похожими друг на друга, имеющими примерно одинаковые свойства и признаки. При этом решение получается на основе анализа значений атрибутов (признаков).

Классификация является одной из важнейших задач *Data Mining*. Она применяется в маркетинге при оценке кредитоспособности заемщиков, определении лояльности клиентов, распознавании образов, медицинской диагностике и многих других приложениях. Если аналитику известны свойства объектов каждого класса, то, когда новое наблюдение относится к

определенному классу, данные свойства автоматически распространяются и на него.

Если число классов ограничено двумя, то имеет место бинарная классификация, к которой могут быть сведены многие более сложные задачи. Например, вместо определения таких степеней кредитного риска, как «Высокий», «Средний» или «Низкий», можно использовать всего две - «Выдать» или «Отказать».

Для классификации в *Data Mining* используется множество различных моделей: нейронные сети, деревья решений, машины опорных векторов, метод k-ближайших соседей, алгоритмы покрытия и др., при построении которых применяется обучение с учителем, когда выходная переменная (метка класса) задана для каждого наблюдения. Формально классификация производится на основе разбиения пространства признаков на области, в пределах каждой из которых многомерные векторы рассматриваются как идентичные. Иными словами, если объект попал в область пространства, ассоциированную с определенным классом, он к нему и относится.

Рассмотрим классификацию данных на примере Ирисов Фишера. Ирисы Фишера - это набор данных для задачи классификации, на примере которого Рональд Фишер в 1936 году продемонстрировал работу разработанного им метода дискриминантного анализа. Этот набор данных стал уже классическим, и часто используется в литературе для иллюстрации работы различных статистических алгоритмов. Ирисы Фишера состоят из данных о 150 экземплярах ириса, по 50 экземпляров из трёх видов - Ирис щетинистый (*Iris setosa*), Ирис виргинский (*Iris virginica*) и Ирис разноцветный (*Iris versicolor*). Для каждого экземпляра измерялись четыре характеристики (в сантиметрах):

- длина чашелистика (англ. *sepal length*);
- ширина чашелистика (англ. *sepal width*);
- длина лепестка (англ. *petal length*);
- ширина лепестка (англ. *petal width*).

На основании этого набора данных требуется построить правило классификации, определяющее вид растения по данным измерений. Это задача многоклассовой классификации, так как имеется три класса - три вида ириса.

2.1 Классификация данных с использованием алгоритма «*g-mean*»

Импортируем в программу данные из файла «Ирисы.txt». Для начала попробуем провести классификацию ирисов, встроенным методом кластеризации *g-mean* (рис. 2.1). Интересно то, что обучение будет проходить без учителя, т.е. выходные данные не будут указаны.

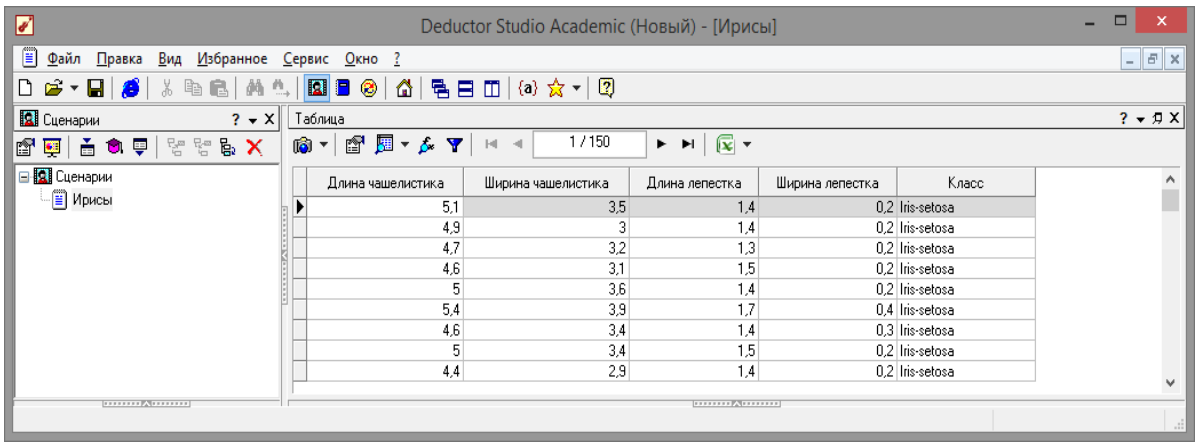


Рис. 2.1 - Импортированные данные

Более того, не будут даже указано количество кластеров, на которое необходимо разделить данную выборку, проверим эффективность встроенной системы кластеризации, для этого необходимо выбрать в мастере обработок пункт кластеризация (рис. 2.2).

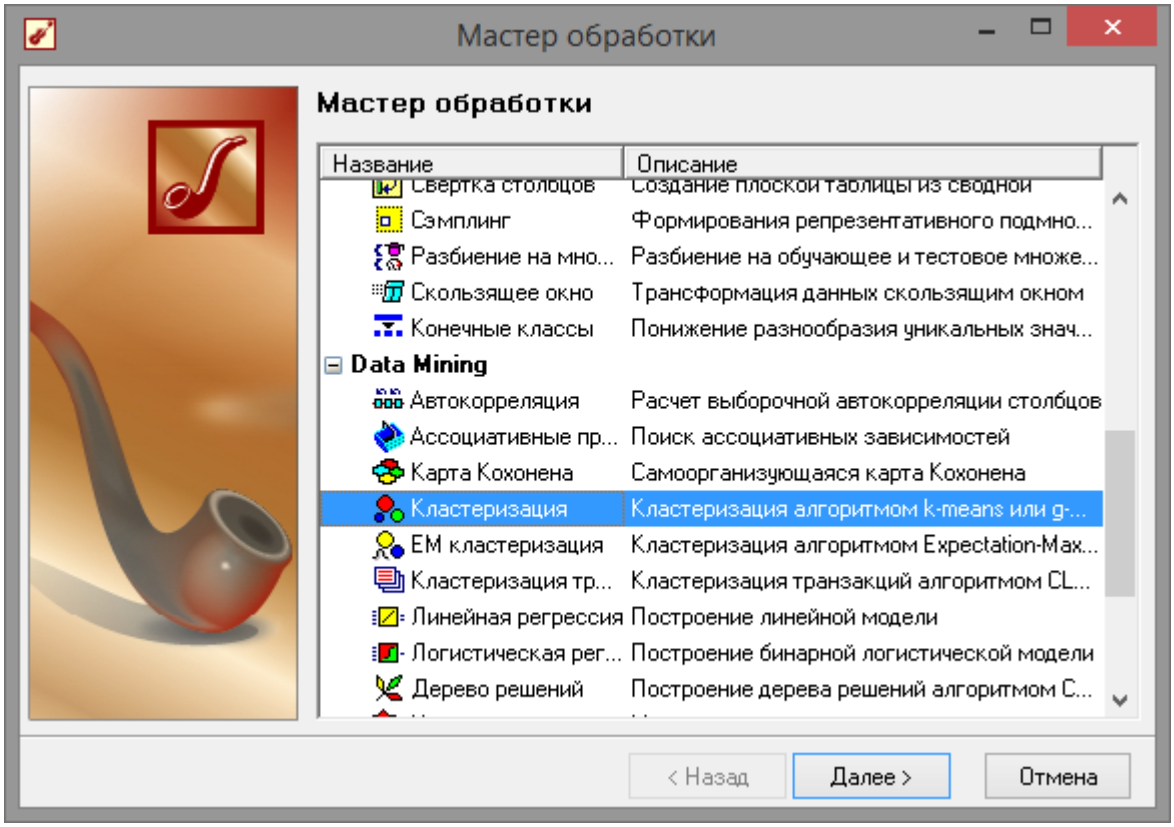


Рис. 2.2 - Мастер кластеризации

На следующем шаге (рис. 2.3) со столбца «Класс» уберем назначение выходное и поставим информационное, теперь обучение будет происходит без учителя.

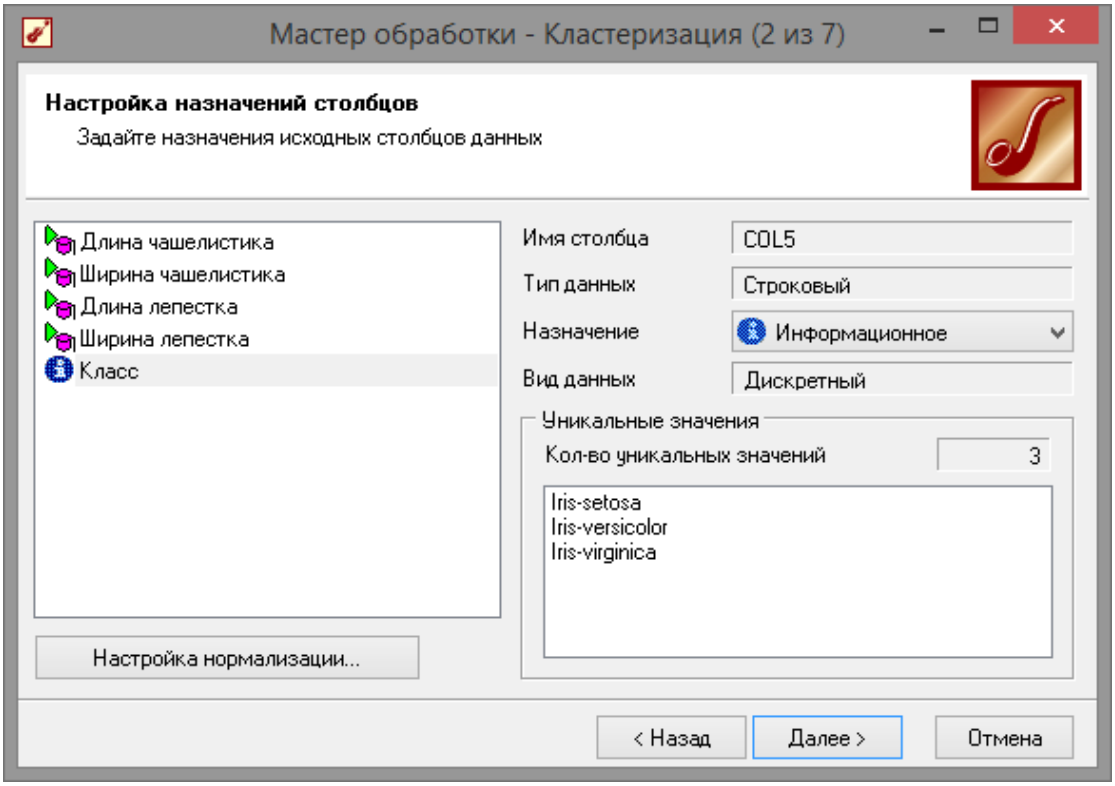


Рис. 2.3 - Настройка мастера кластеризации

Далее следует настроить параметры обучения, в данном конкретном случае, параметры по умолчанию отлично подходят (рис. 2.4).

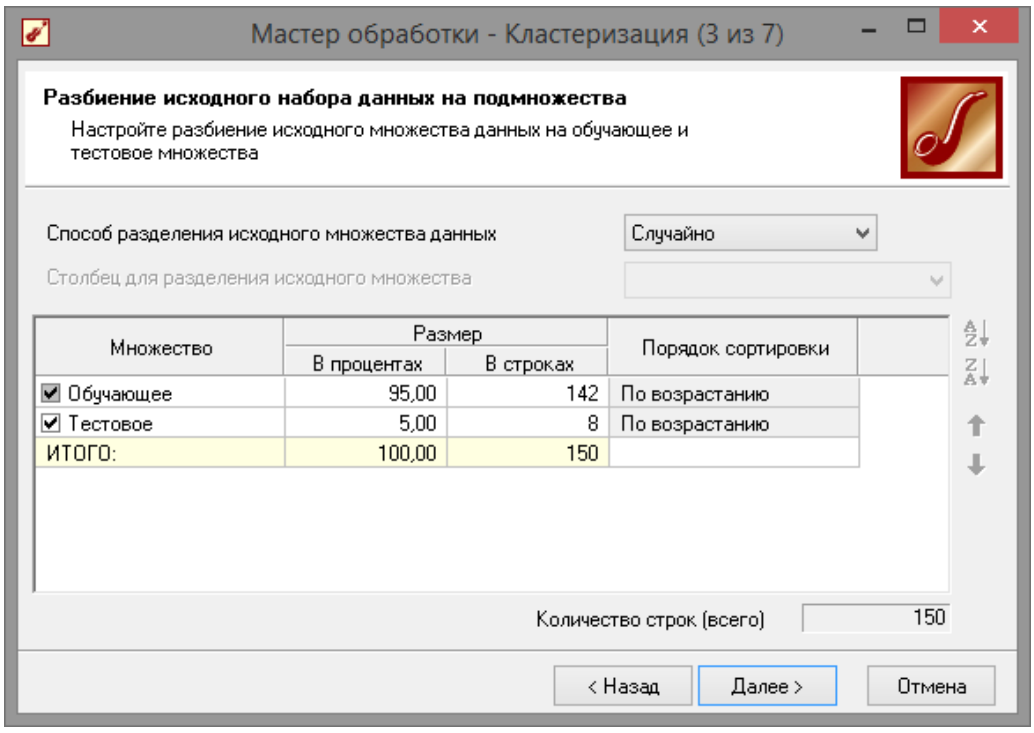


Рис. 2.4 - Настройка параметров обучения

На следующем шаге мастера необходимо выбрать алгоритм

кластеризации. На и так заранее известно, что количество кластеров должно равняться «3», но протестируем возможности программы и предоставим ей самой выбрать количество кластеров. Это так же будет полезно, если не известно на какое количество групп следует разбивать выборку (рис. 2.5).

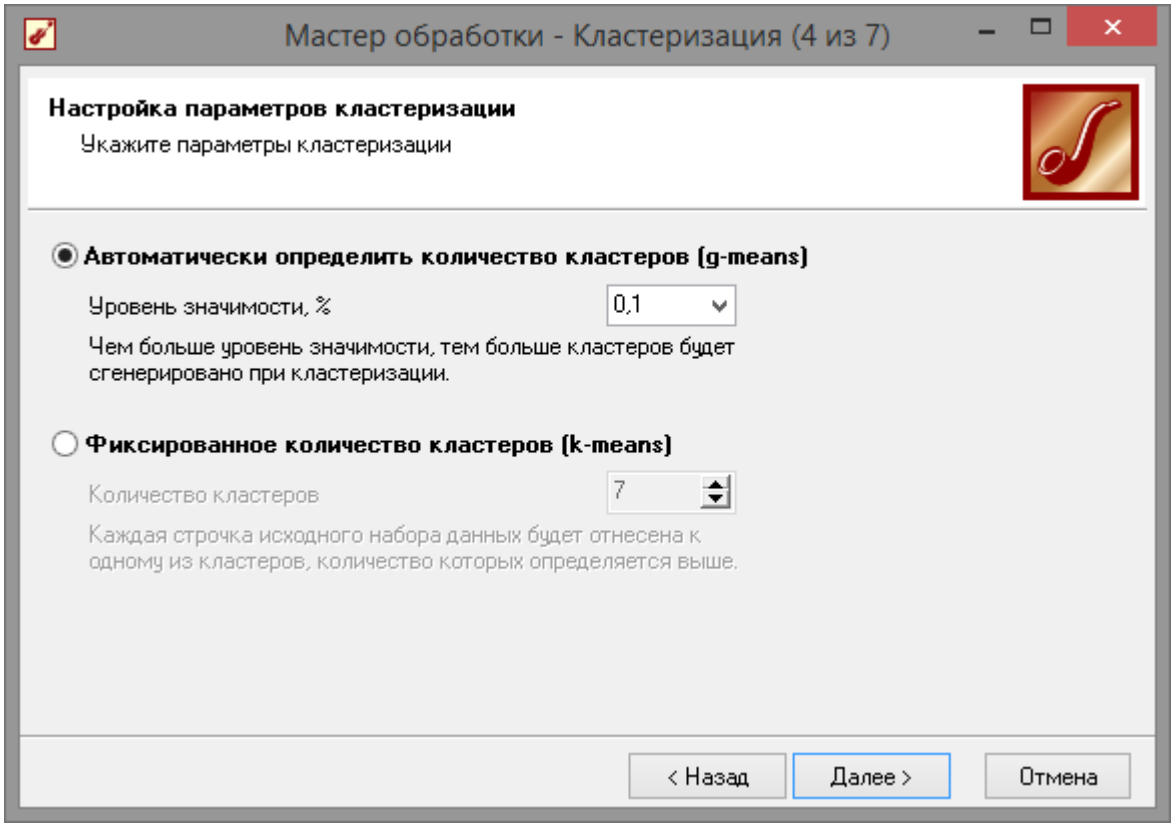


Рис. 2.5 - Выбор алгоритма кластеризации

Следующий шаг мастера предлагает запустить процесс обучения и наблюдать в процессе обучения величину ошибки, а также процент распознанных примеров. Параметр «Частота обновления» отвечает за то, через какое количество эпох обучения выводится данная информация (рис. 2.6).

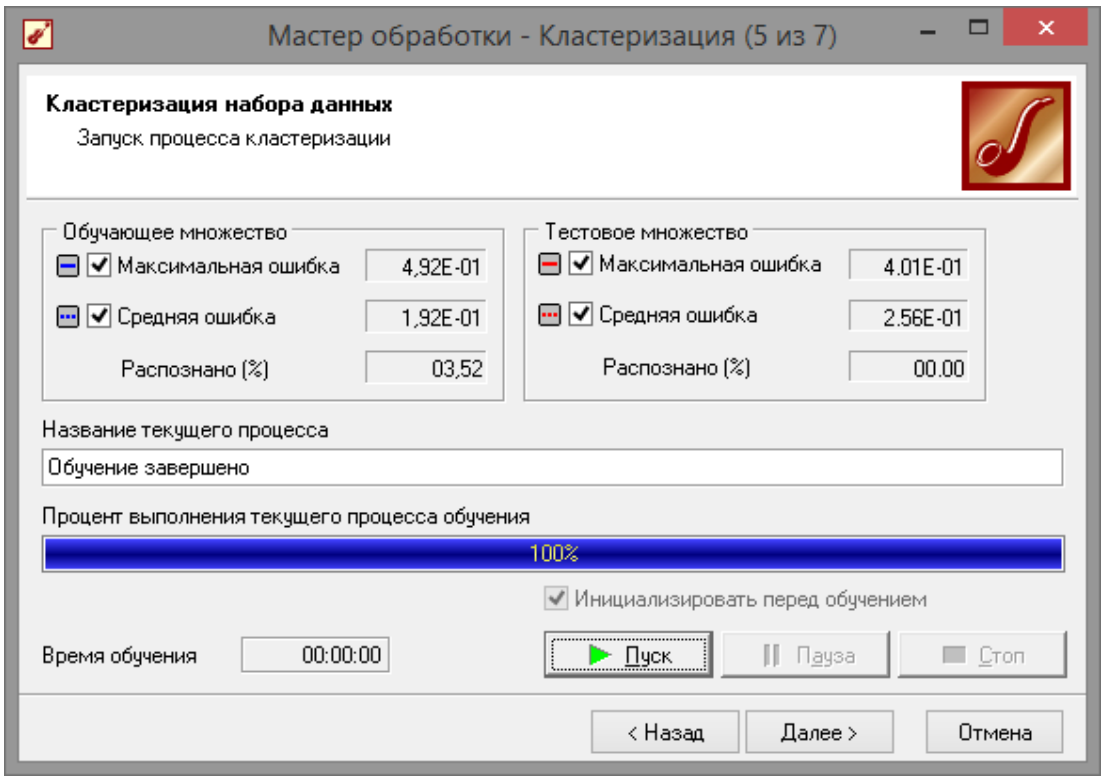


Рис. 2.6 - Обучение сети

После обучения сети, в качестве визуализаторов выберем: «Связи кластеров»; «Профили кластеров»; «Матрицу сравнения»; «Что-если» (рис. 2.7).

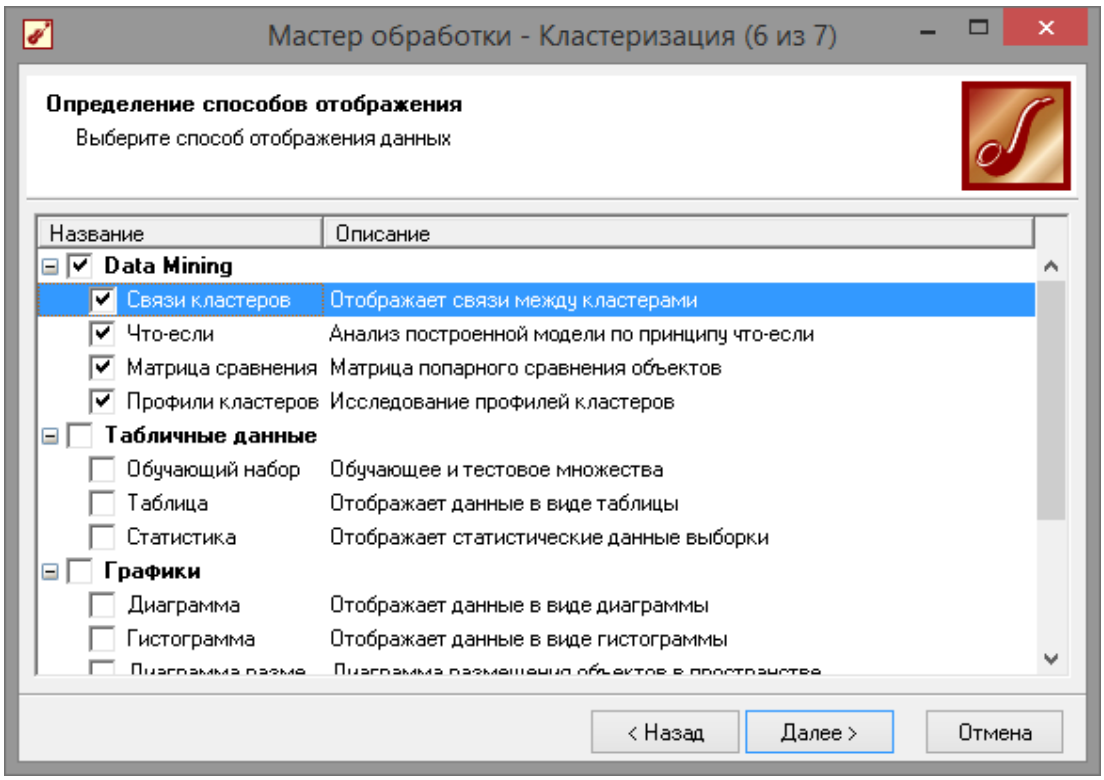


Рис. 2.7 - Визуализация данных

Как видно из рис. 2.8 приложение верно переделило количество групп, также здесь можно увидеть на сколько значим тот или иной параметр, для присвоения цветку того или иного вида.

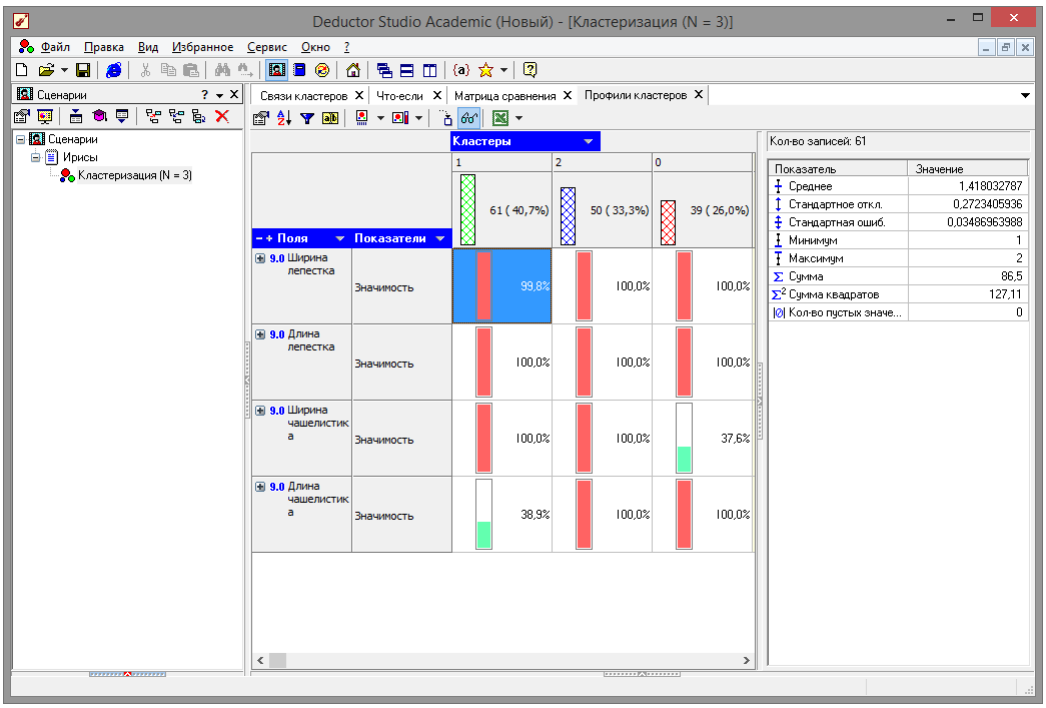


Рис. 2.8 - Профили кластеров

На матрице сравнения (рис. 2.9) видно, что больше всех от остальных отличается «кластер 2» (ему соответствует «*iris-setosa*»).

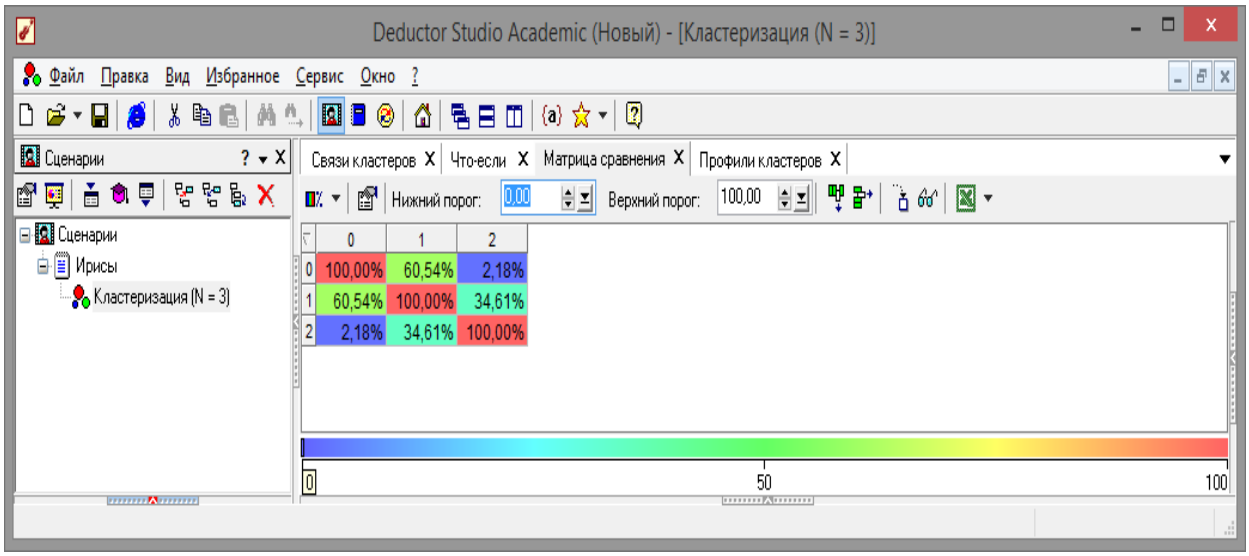


Рис. 2.9 - Матрица сравнения

Визуализатор «Что-если» (рис. 2.10) позволит провести эксперимент, введя любые значения параметров. Если же нажать кнопку «Загрузить данные» из исходной выборки, то можно заметить неточности определения кластеров из-за схожих параметров цветков.

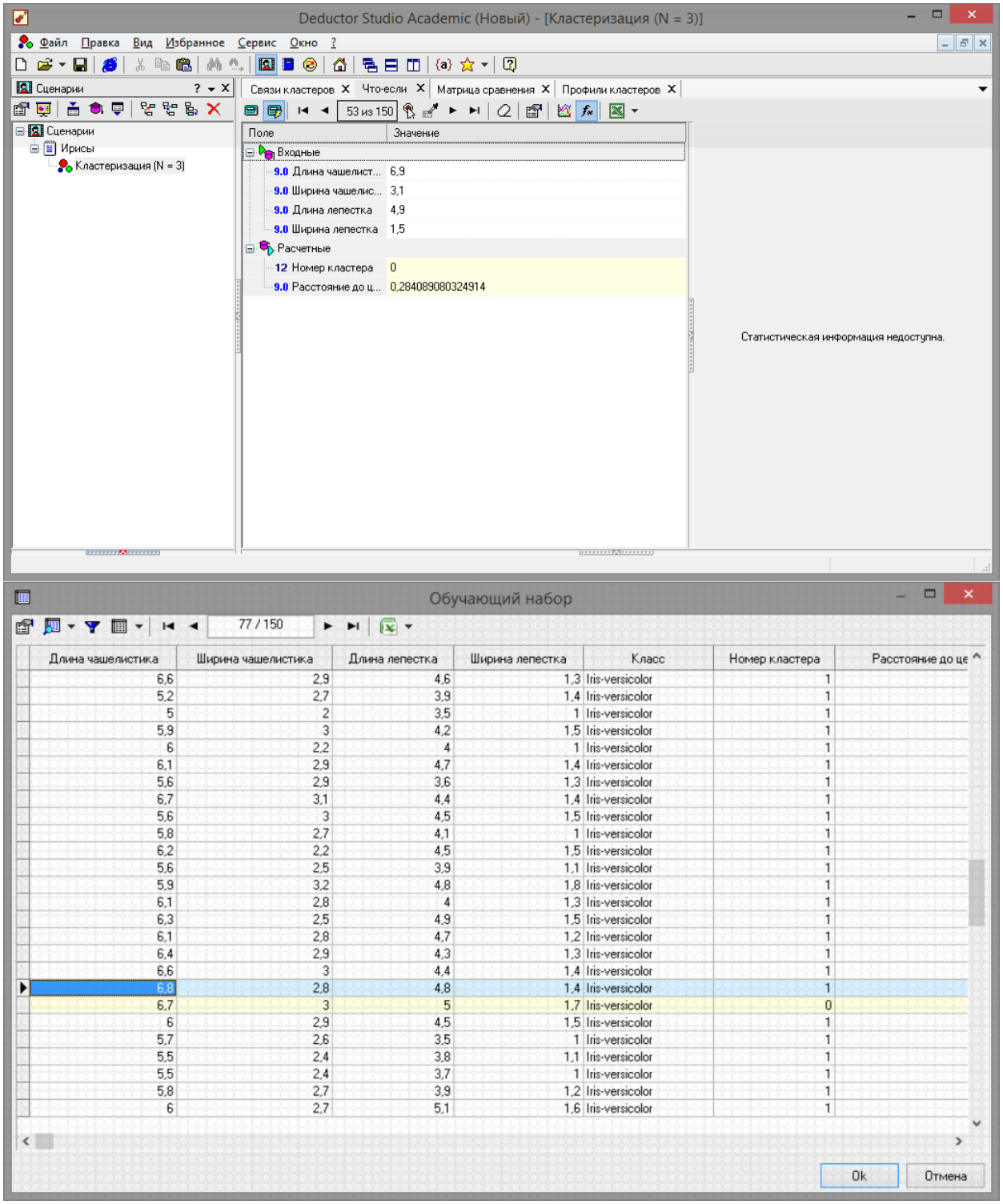


Рис. 2.10 - Инструмент «Что-Если»

2.2 Классификация данных с помощью нейронной сети

Теперь опробуем провести классификацию ирисов при помощи обычной нейронной сети. Входные данные брать из файла

«Ирисы.txt», только в данном случае обучение будет происходить с учителем. Импортируем данные и в мастере обработок выберем пункт

нейросеть (рис. 2.11).

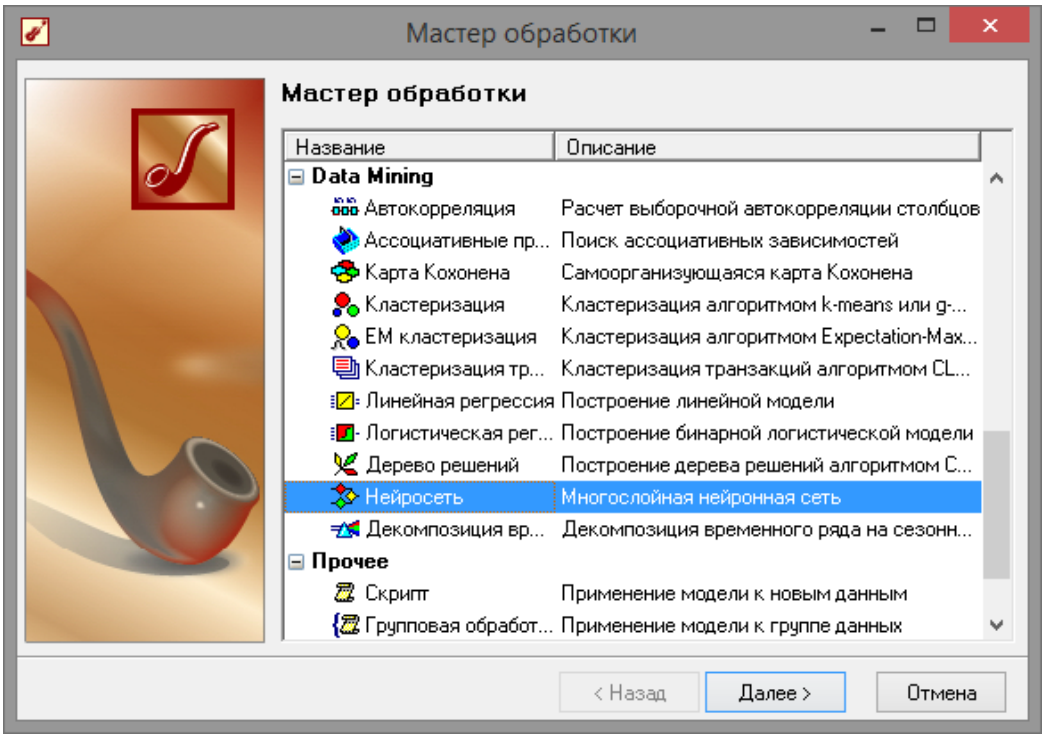


Рис. 2.11 - Мастер обработок

Установим в качестве выходного значения параметр класс. В качестве нормализатора для выхода «Класс» необходимо установить пункт «Уникальные значения».

Примечание. Уникальные значения используются для дискретных значений. Такими являются строки, числа или даты, заданные дискретно. Чтобы привести непрерывные числа в дискретные, можно, например, воспользоваться обработкой

«Квантование». Так следует поступать для величин, для которых можно задать отношение порядка, то есть, если для двух любых дискретных значений можно указать, какое больше, а какое меньше. Тогда все значения необходимо расположить в порядке возрастания (рис. 2.12-2.13). Далее они нумеруются по порядку, и значения заменяются их порядковым номером

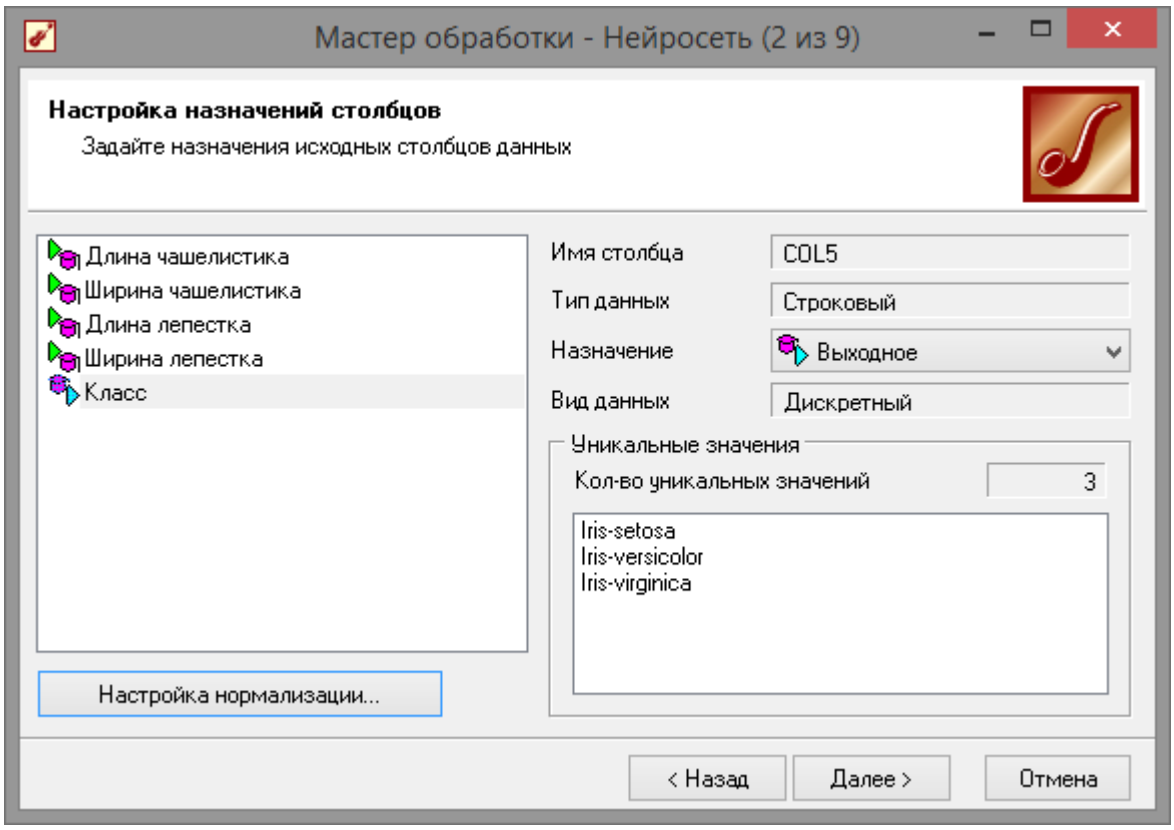


Рис. 2.12 - Назначение входов и выходов

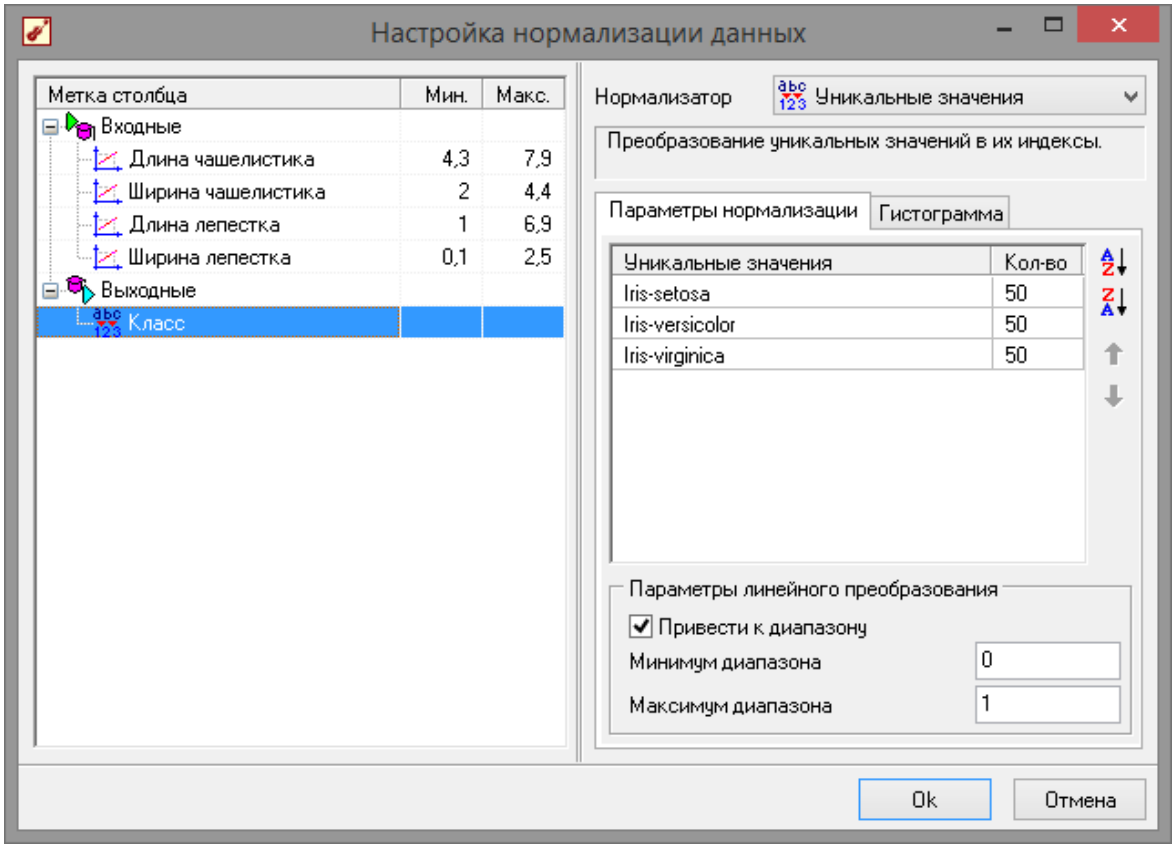


Рис. 2.13 - Настройка нормализации

Следующим шагом указываем параметры обучения (рис. 2.14).

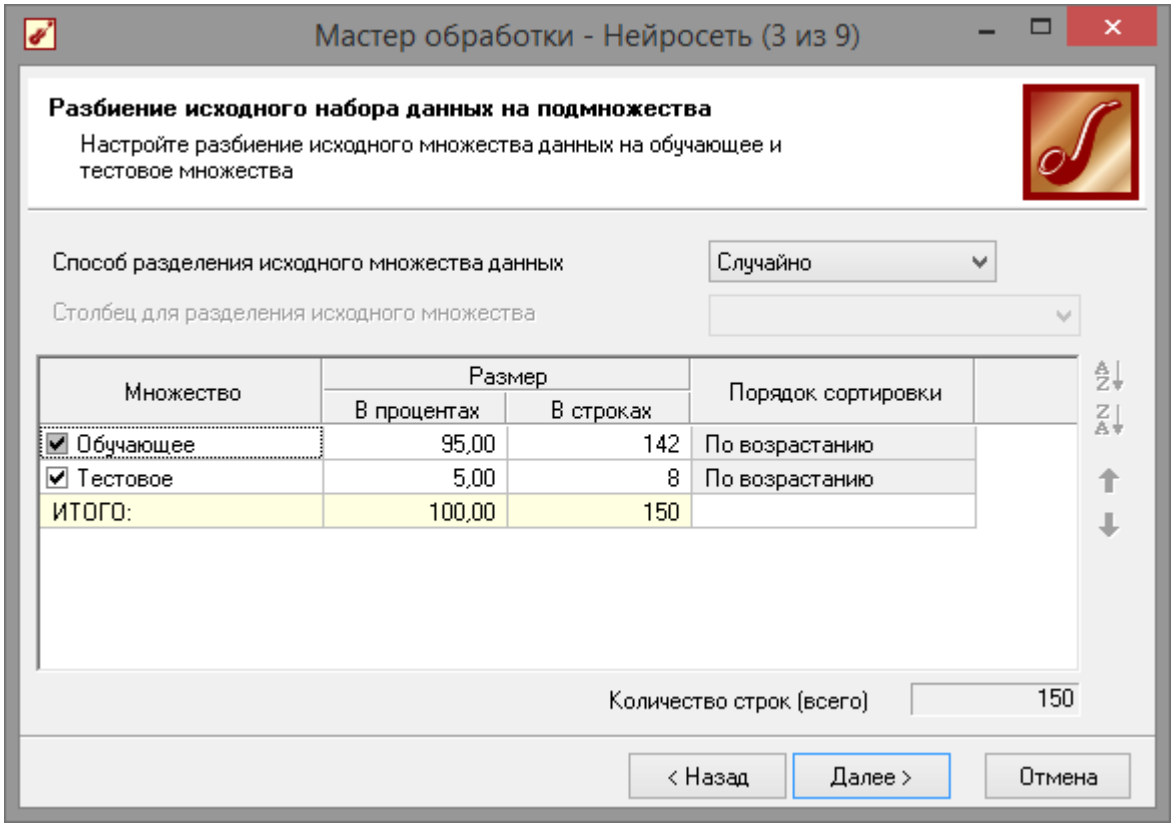


Рис. 2.14 - Параметры обучения

Количество нейронов первого слоя экспериментально было выставлено в значение 5 (рис. 2.15).

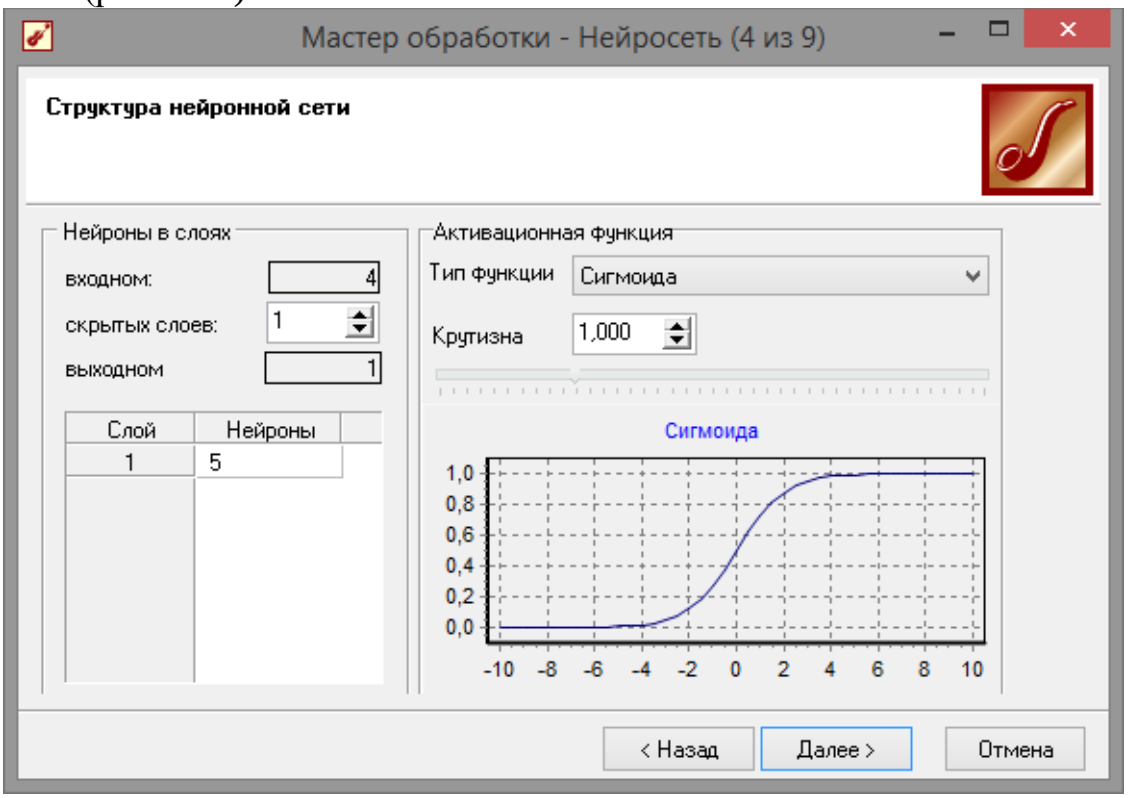


Рис. 2.15 - Параметры обучения

Далее выбираем алгоритм обучения (рис. 2.16).

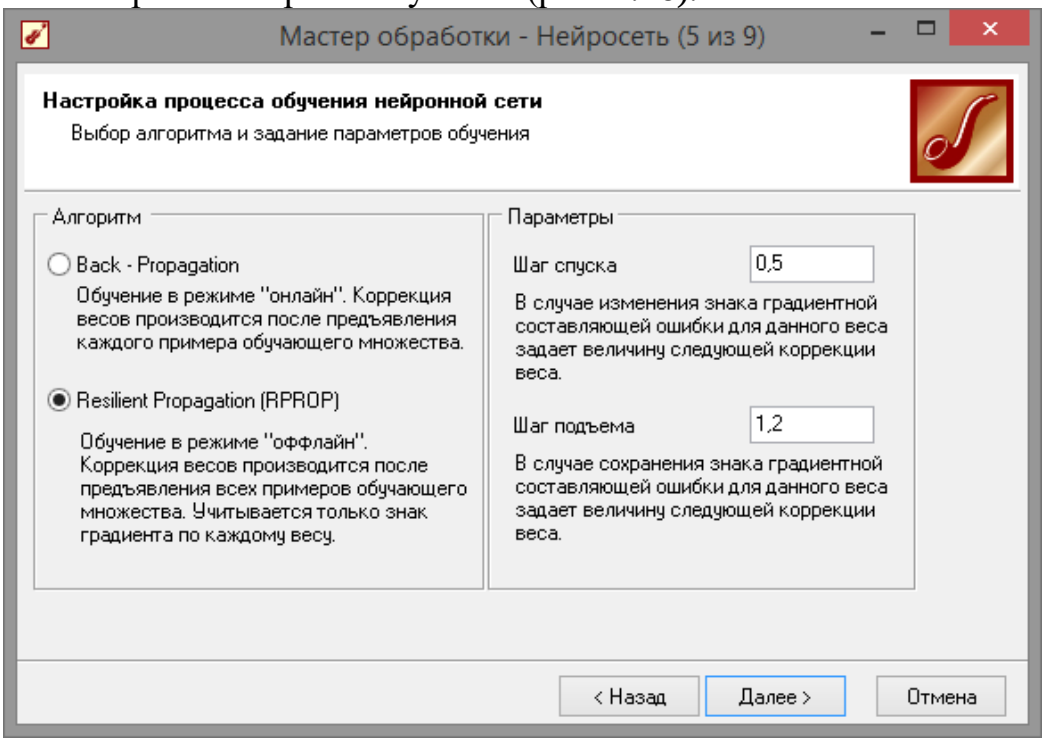


Рис. 2.16 - Алгоритм обучения

На следующем шаге (рис. 2.17) устанавливаем значение погрешности и количество эпох обучения. И запускаем обучение сети (рис. 2.18).

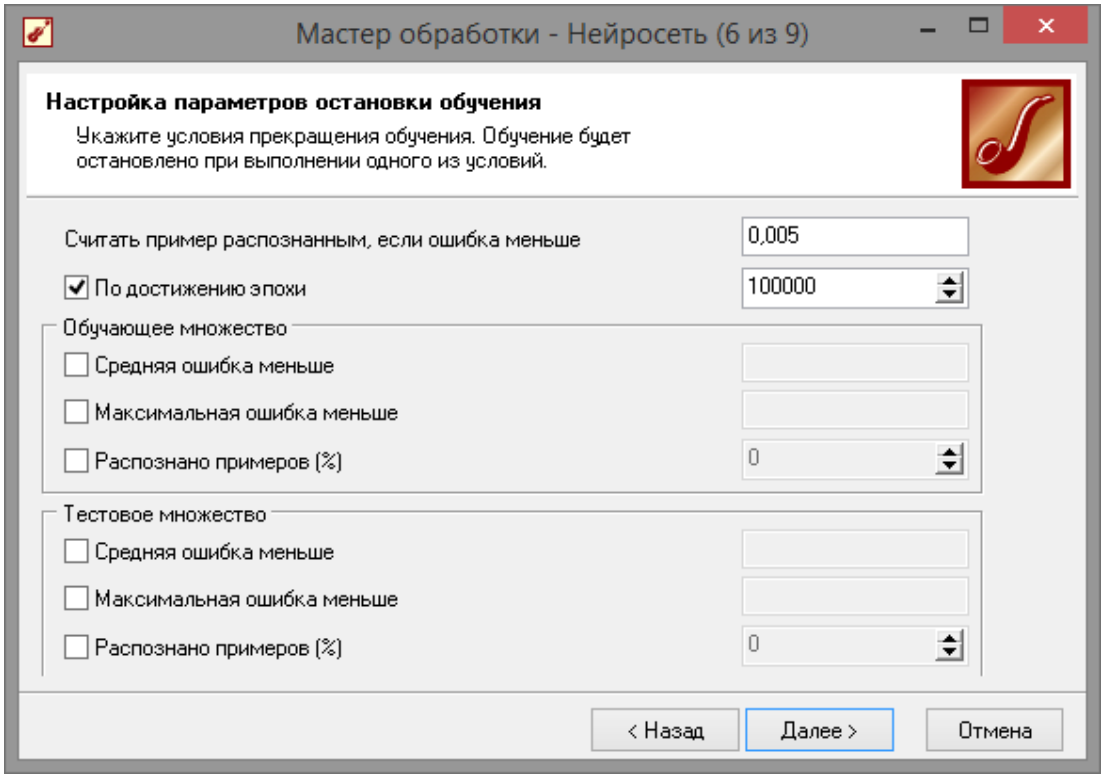


Рис. 2.17 - Установка погрешности и числа эпох обучения

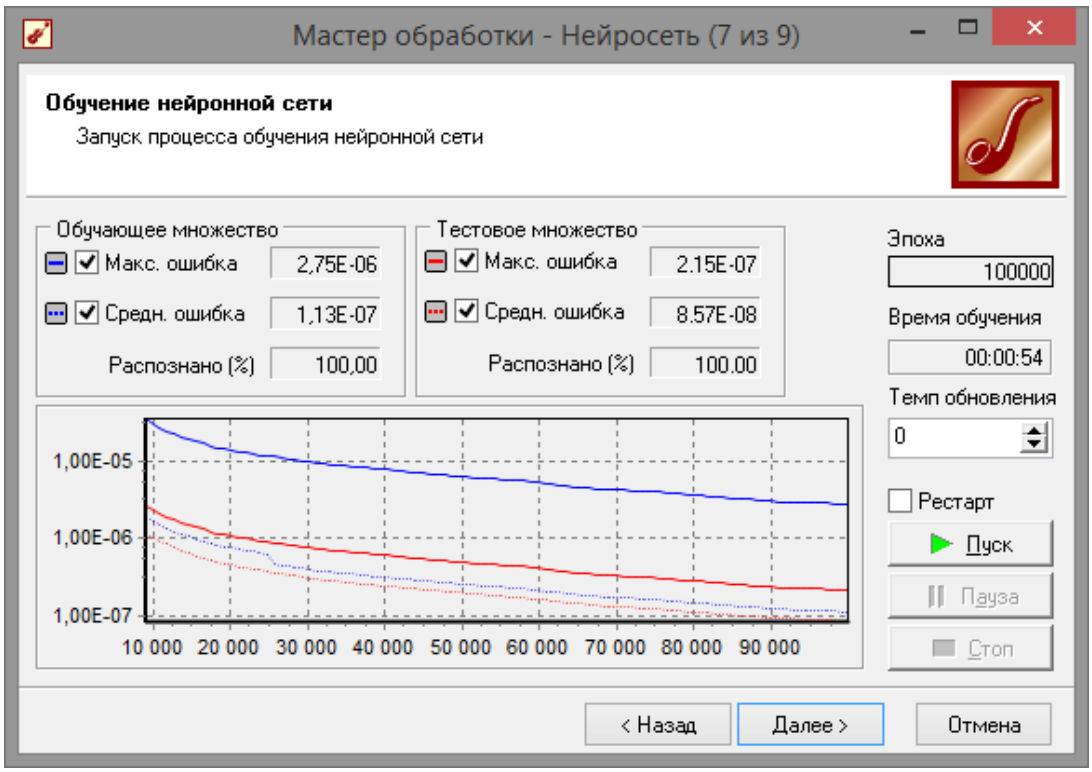


Рис. 2.18 - Обучение сети

Как видно из рис. 2.18, сеть обучалась дольше, но дала более точные результаты. В качестве визуализаторов выбираем «Граф нейросети», далее пункт «Таблица сопряжения» и «Что-Если» (рис. 2.19).

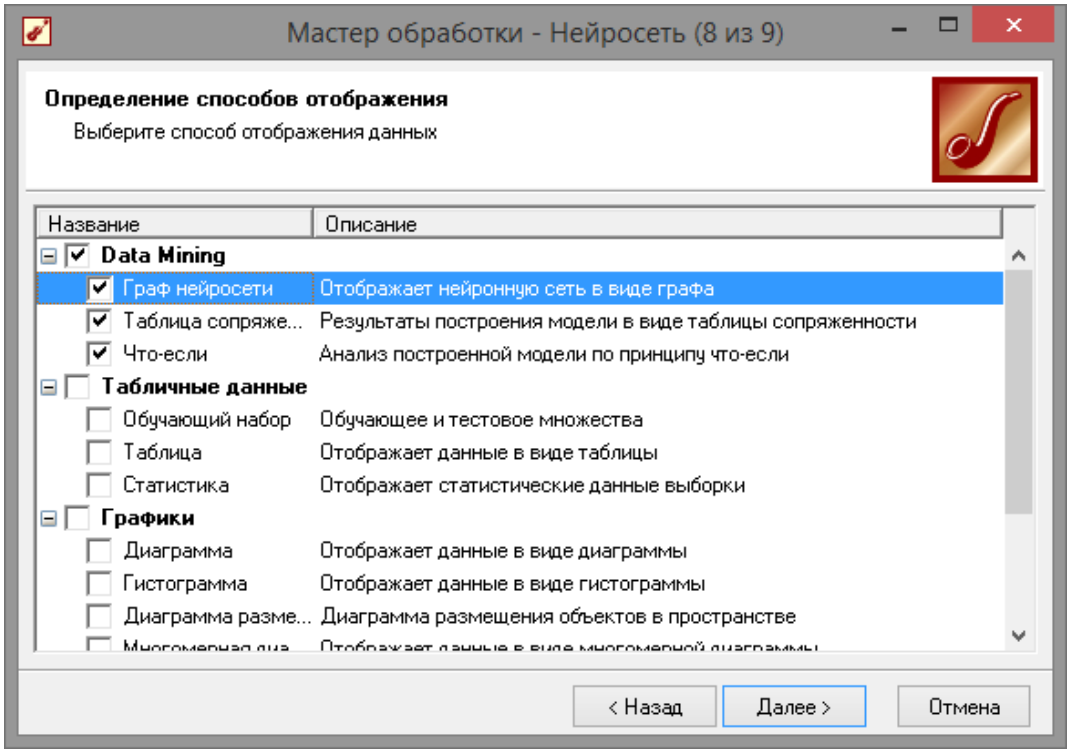


Рис. 2.19 – Визуализаторы

На графе нейросети видно, как выглядит обученная сеть (рис. 2.20).

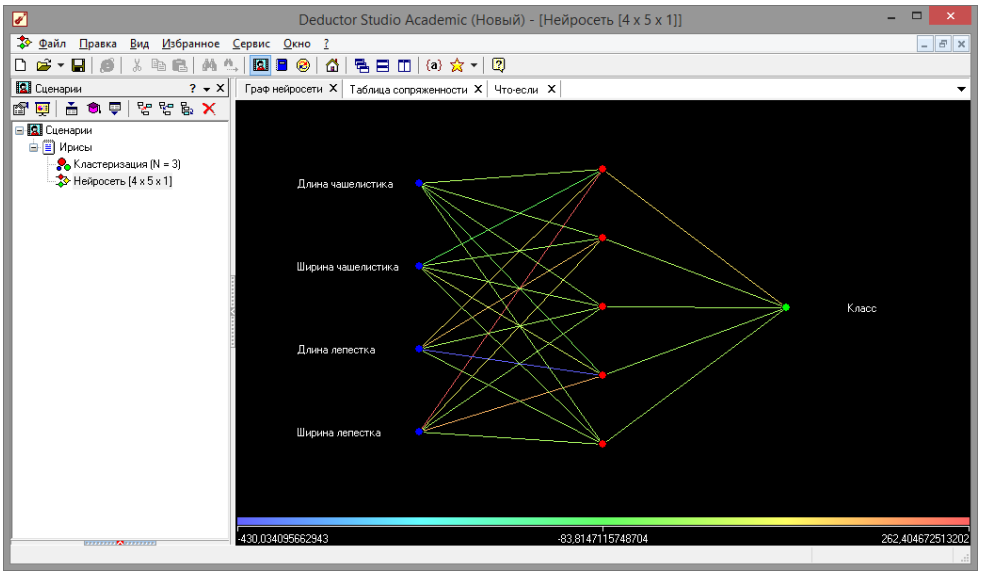


Рис. 2.20 - Граф нейросети

Диаграмма сопряженности (рис. 2.21) показывает, как распределились входные значения. На ней видно что сеть обучилась абсолютно точно.

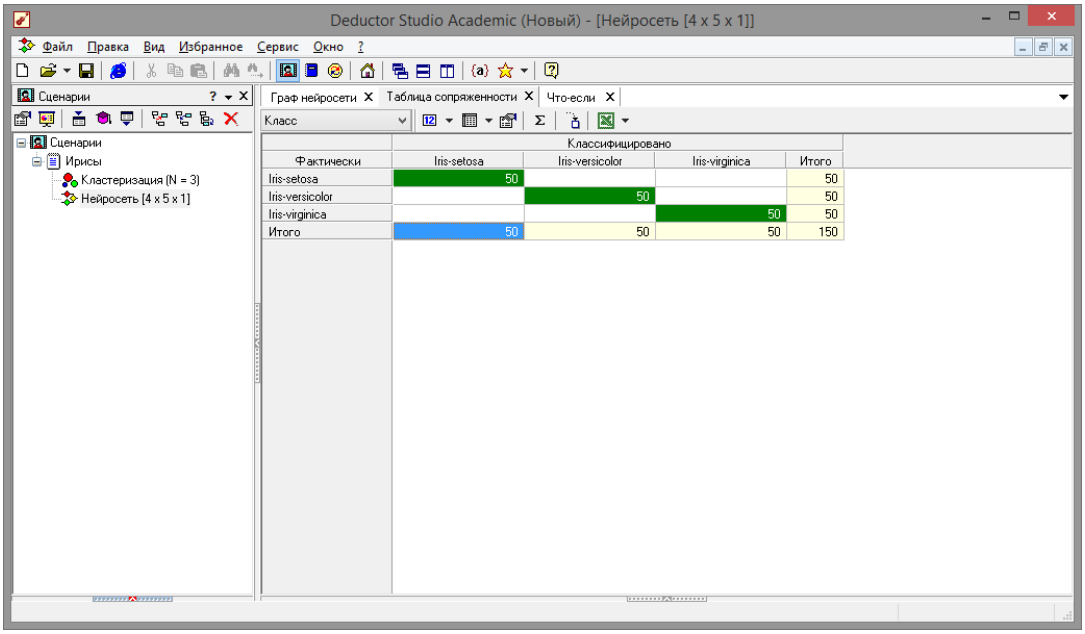
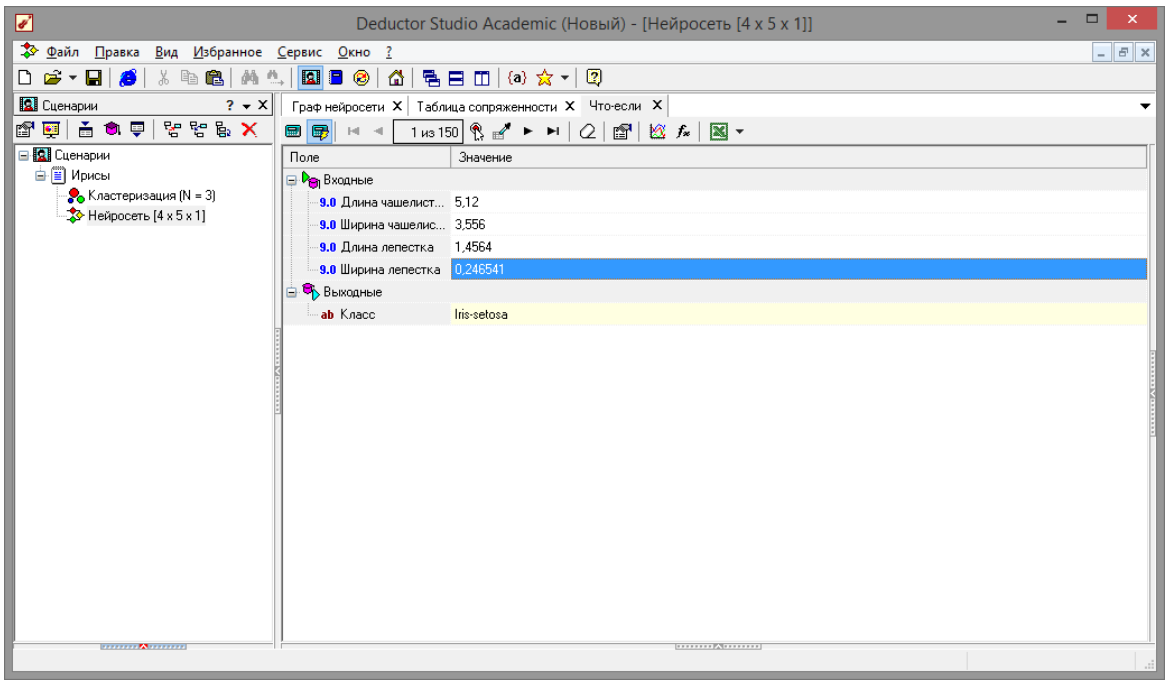


Рис. 2.21 - Диаграмма сопряженности

Убедимся в высказанном выше заключении взглянув на таблицу «Что-Если» (рис. 2.22). Даже при вводе значений отсутствующих в выборке, сеть верно реагирует на них. Далее можно опять нажать на

кнопку «Загрузить данные из исходной выборки» чтобы убедиться в правильности распознавания результатов.



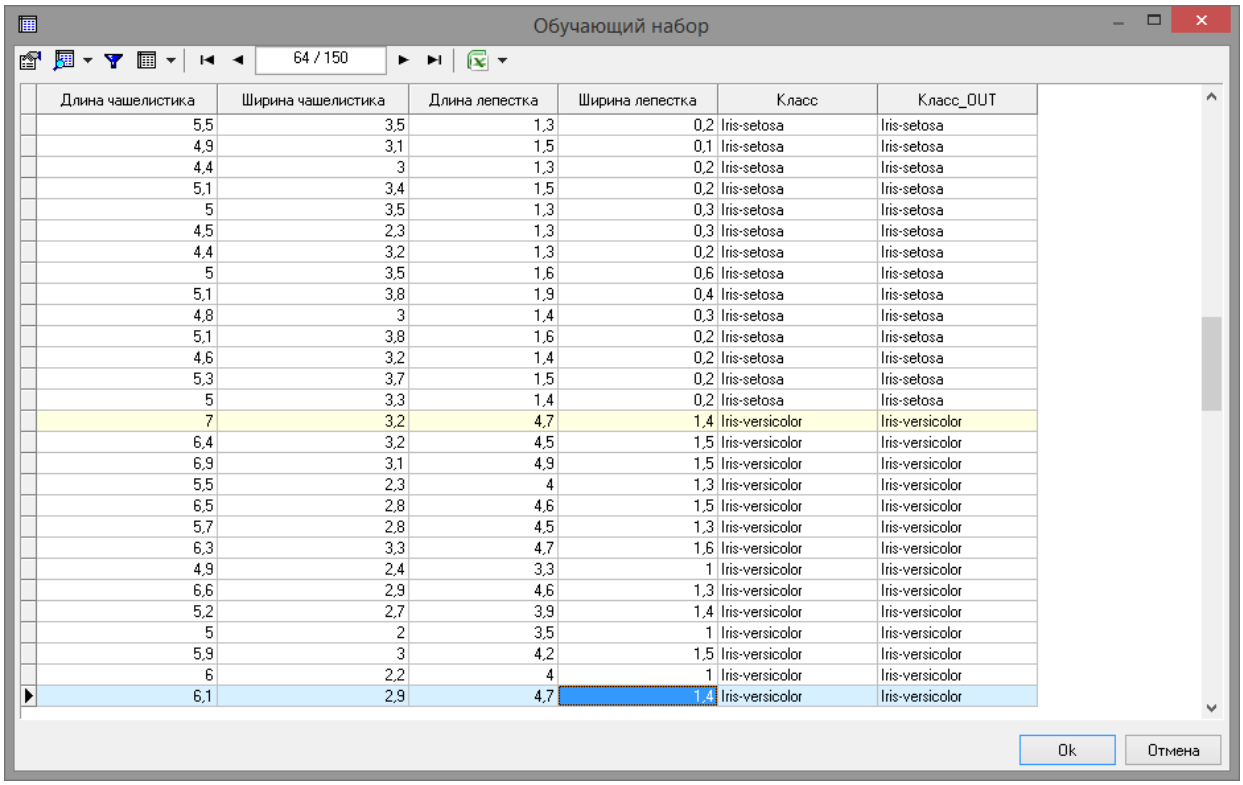


Рис. 2.22 - Инструмент «Что-Если»

Данный пример показал, как можно обучить сеть с учителем. Был изучен новый инструмент «Нейросеть» и сделано предположение, что обучение с учителем более эффективно, хотя и занимает больший промежуток времени.

Пример показал простоту и удобство применения «Нейросеть» для классификации Ирисов Фишера. Мастер предлагает широкие возможности по настройке процесса обучения. После обучения сети стали видны ее

достоинства для анализа. Также были продемонстрированы широкие возможности визуализации простроенной сети. Все это говорит о эффективности инструмента «Нейросеть».

2.3 Задания

1. Сгенерировать данные для классификации, провести классификацию двумя способами (инструмент «Кластеризация» и «Нейронная сеть»), сделать выводы по эффективности этих двух способов.
2. Провести эксперимент по изменению параметров обучения нейросети, и сделать выводы по эффективности процесса обучения при разном количестве нейронов.

Контрольные вопросы

1. Для чего служат алгоритмы g-mean и k-mean?
2. Какие алгоритмы обучения нейронной сети предлагает программный комплекс Deductor Academic?
3. В чем их отличие?
4. Что такое обучение с учителем?
5. Что такое обучение без учителя?

Лабораторная работа 3

Применение интеллектуального анализа данных в задачах поддержки принятия решений

Цель работы: освоить методы и средства прогнозирования в пакете *Deductor Academic* при интеллектуальном анализе данных в задачах поддержки принятия решений.

Программа работы

1. Выполнить пример прогнозирования с помощью нейронных сетей в пакете Deductor Academic.
2. Выполнить пример прогнозирования с помощью временных рядов в пакете Deductor Academic.
3. Выполнить прогнозирование с помощью нейронных сетей и временных рядов на данных согласно индивидуальному заданию.

Методические указания по выполнению работы

Основное направление программы *Deductor Studo* – анализ, прогнозирование, классификация и кластеризация данных. Программа предоставляет следующие механизмы анализа: нейронные сети, линейный регрессионный анализ, построение деревьев решений, самоорганизующиеся карты Кохонена, прогнозирование временного ряда, обнаружение дубликатов и противоречий. Нейросети – механизм, который используют для прогнозирования и решения задач классификации. Они применяются в основном там, где существует нелинейные зависимости результата от входных факторов.

3.1 Прогнозирование умножения с помощью нейронных сетей

Рассмотрим прогнозирование с помощью нейронных сетей на примере прогнозирования результата умножения двух чисел – файл

«Произведение.txt». В нем содержится таблица со следующими полями: «АРГУМЕНТ1», «АРГУМЕНТ2» – множители,

«ПРОИЗВЕДЕНИЕ» – их произведение. Причем произведение некоторых чисел пропущено в обучающей выборке, например, 7x7=49. Импортировав данные из файла, можно посмотреть результат умножения, используя таблицу.

Пусть необходимо построить модель прогноза умножения, подавая на вход которой два множителя получать на выходе их произведение. Для этого необходимо, находясь на узле импорта, открыть мастер обработки, показанный на рис. 3.1. В нем выбрать в качестве обработки нейронную сеть и перейти к следующему шагу мастера. На втором шаге мастера необходимо установить назначение полей «АРГУМЕНТ1» и «АРГУМЕНТ2» как входные, а поле

«ПРОИЗВЕДЕНИЕ» – как выходное и задать тип данных как

«Целый» (рис. 3.2).

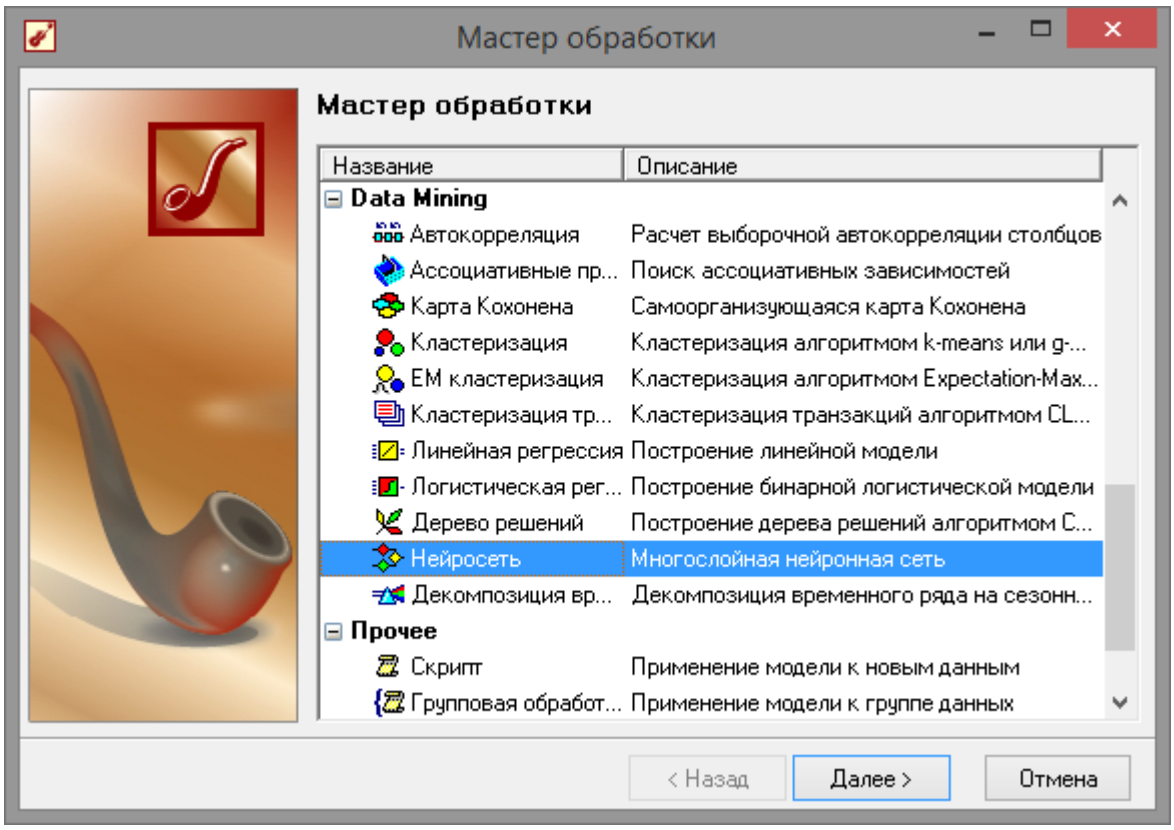


Рис. 3.1 - Мастер обработки

На следующем шаге предлагается настроить разбиение исходного множества данных на обучающее и тестовое. Здесь необходимо указать способ разбиения исходного множества данных

«Случайно» и поставить обучающее множество размером 100 %, так как выборка слишком мала (рис. 3.3).

Далее необходимо указать функцию активации, количество скрытых слоев, и количество нейронов в слое. Данные необходимо подбирать экспериментально, так как большое количество нейронов и слоев, может сильно замедлить процесс обучения. Эмпирически были выставлены настройки, показанные на рис. 3.4.

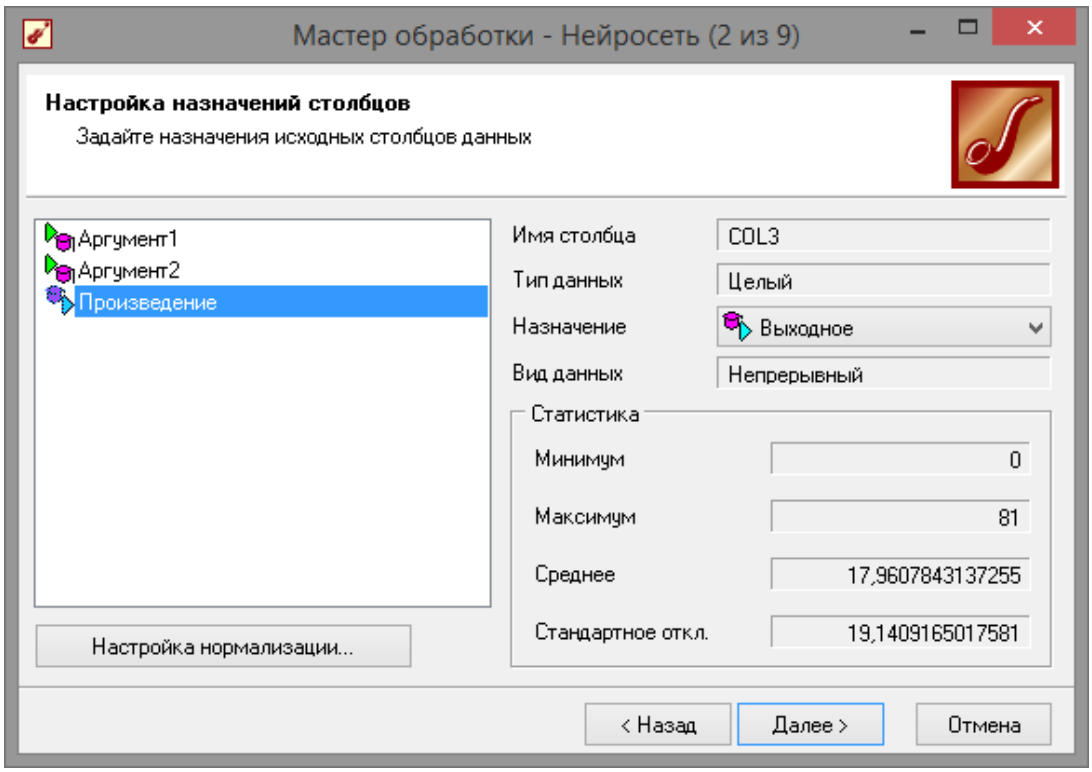


Рис. 3.2 - Мастер нейросети

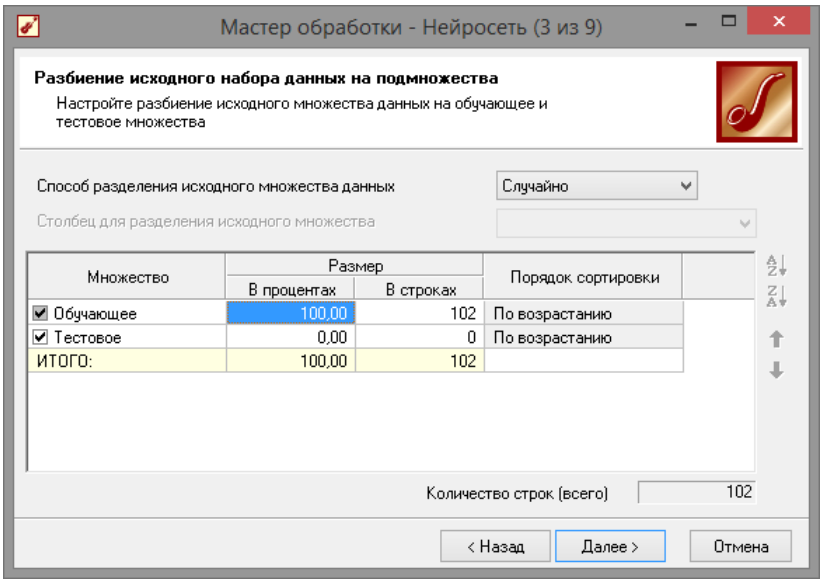


Рис. 3.3 - Настройки параметров обучения

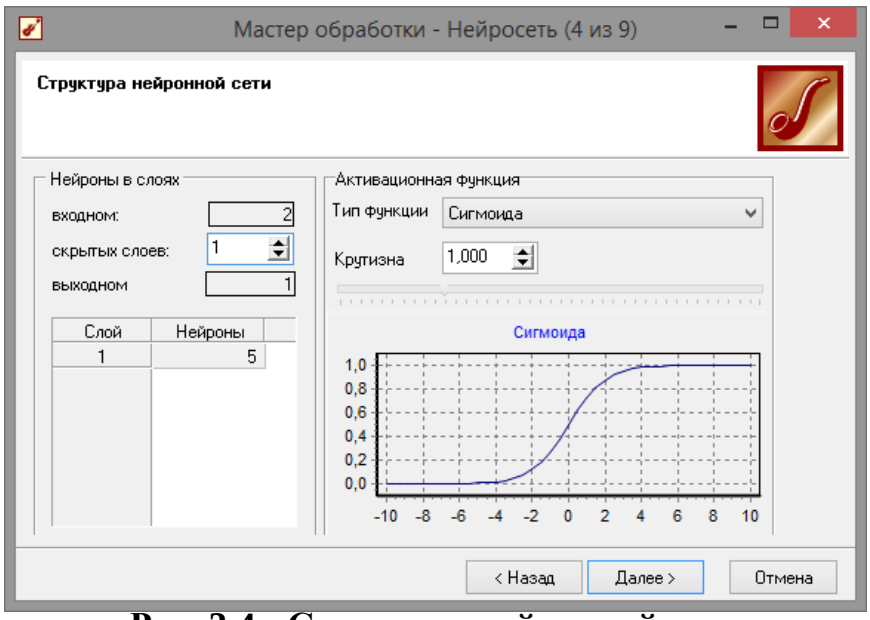


Рис. 3.4 - Структура нейронной сети

Следующий шаг предлагает выбрать алгоритм обучения и его параметры (рис. 3.5).

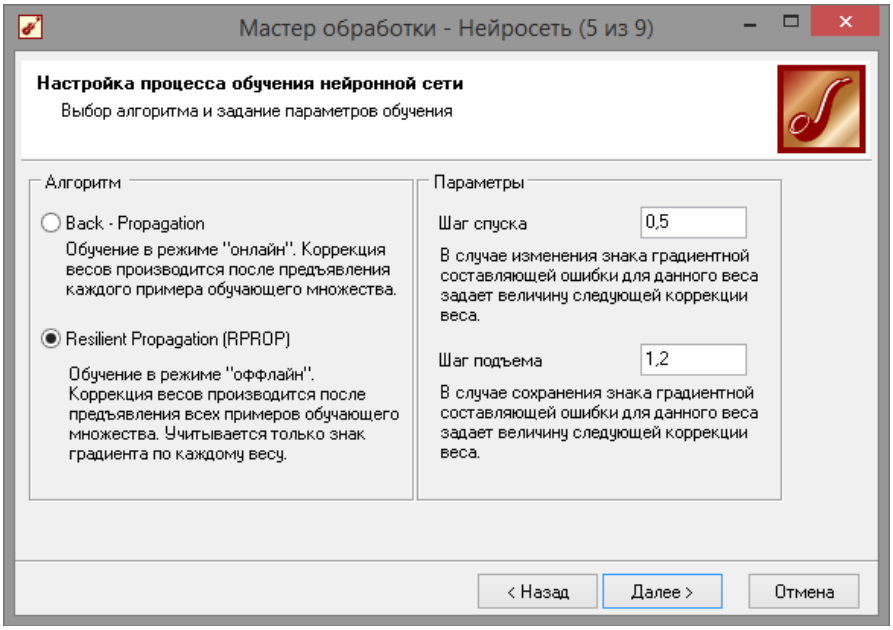


Рис. 3.5 - Алгоритм обучения

Далее настроим условия остановки обучения (рис. 3.6). Пусть пример считаем распознанным, если ошибка меньше 0,005. Также укажем условие остановки обучения при достижении эпохи 100000.

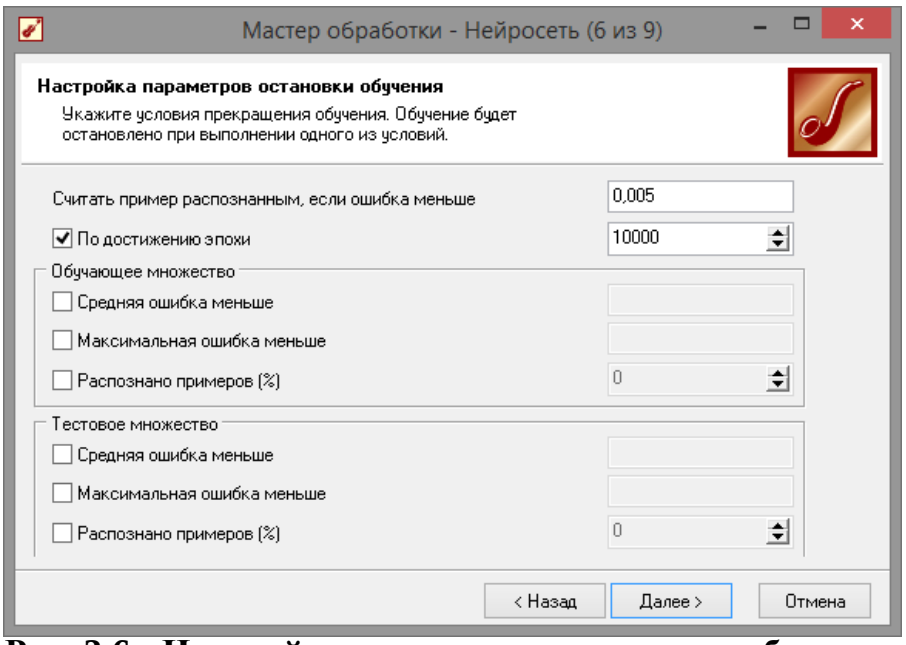


Рис. 3.6 - Настройка параметров остановки обучения

Следующий шаг мастера (рис. 3.7) предлагает запустить процесс обучения и наблюдать в процессе обучения величину ошибки, а также процент распознанных примеров. Параметр «Частота обновления» отвечает за то, через какое количество эпох обучения выводится данная информация.

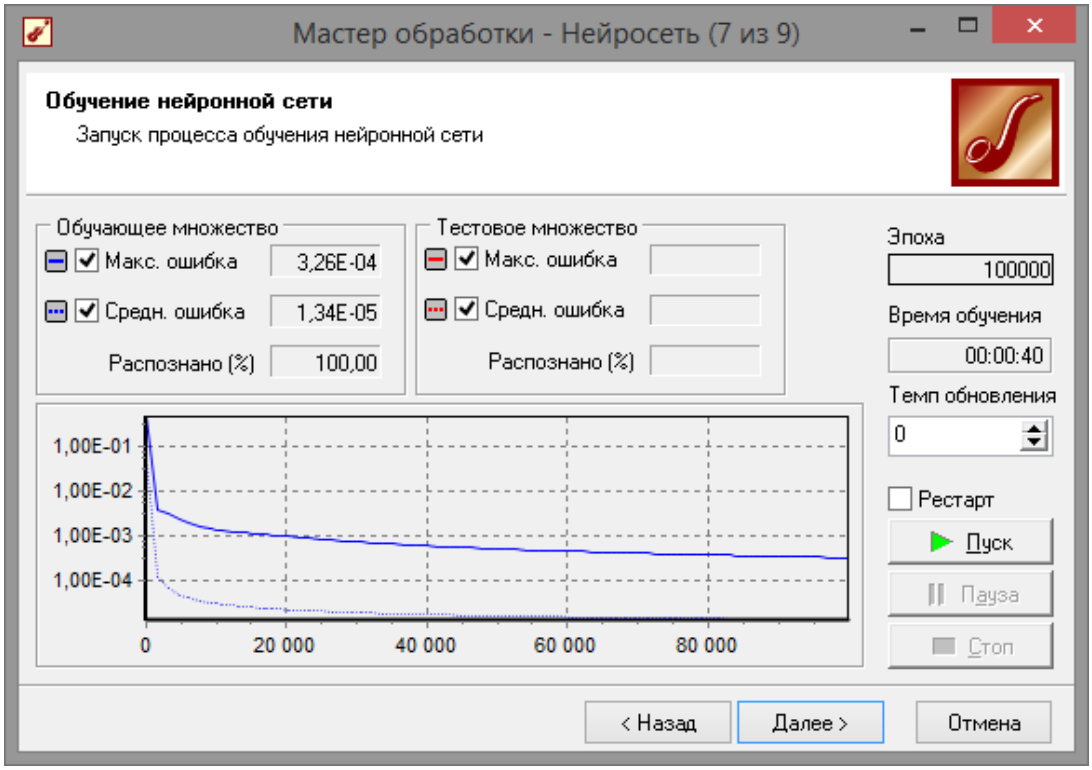


Рис. 3.7 - Обучение сети

После обучения сети, в качестве визуализаторов выберем варианты,

показанные на рис. 3.8: «Диаграмма», «Диаграмма рассеяния», «Граф нейросети», «Что-если».

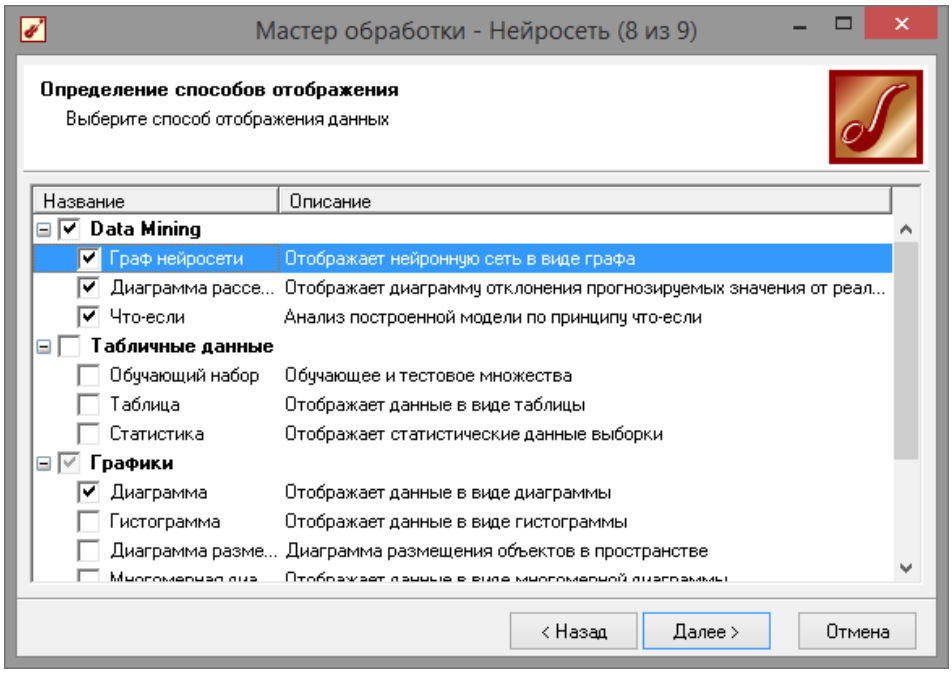


Рис. 3.8 - Визуализация данных

Результаты наглядно видны на диаграмме рассеяния (рис. 3.9), которая показывает рассеяние прогнозируемых данных относительно эталонных. По диаграмме можно судить, что сеть обучена недостаточно хорошо, из-за малого объема обучающей выборки. Также можно сравнить эталонные данные с прогнозируемыми, выбрав на обычной диаграмме два поля – «ПРОИЗВЕДЕНИЕ» и

«ПРОИЗВЕДЕНИЕ_OUT». Если масштабировать диаграмму и включить отображение меток, то можно увидеть достаточно большую ошибку (рис. 3.10), но на цели этого задания она сказываться не будет.

Визуализатор «Что-если» позволит провести эксперимент, введя любые значения множителей АРГУМЕНТ1 и АРГУМЕНТ2 и рассчитав результат их произведения. Попробуем ввести аргументы, которые отсутствуют в обучающей выборке, например, 7x7. Как видно на рис. 3.11 сеть обучена достаточно точно, так как получен верный результат, несмотря на то, что сеть никогда не видела такую комбинацию аргументов.

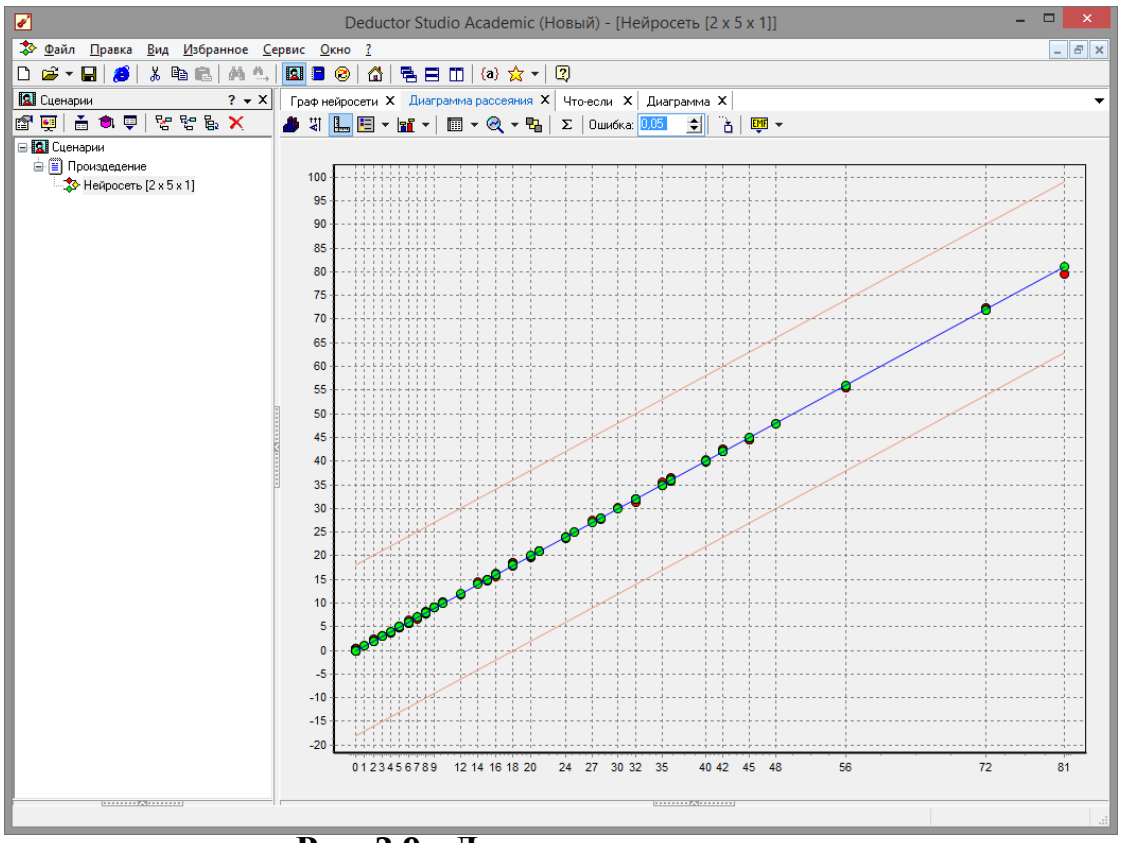


Рис. 3.9 - Диаграмма рассеяния

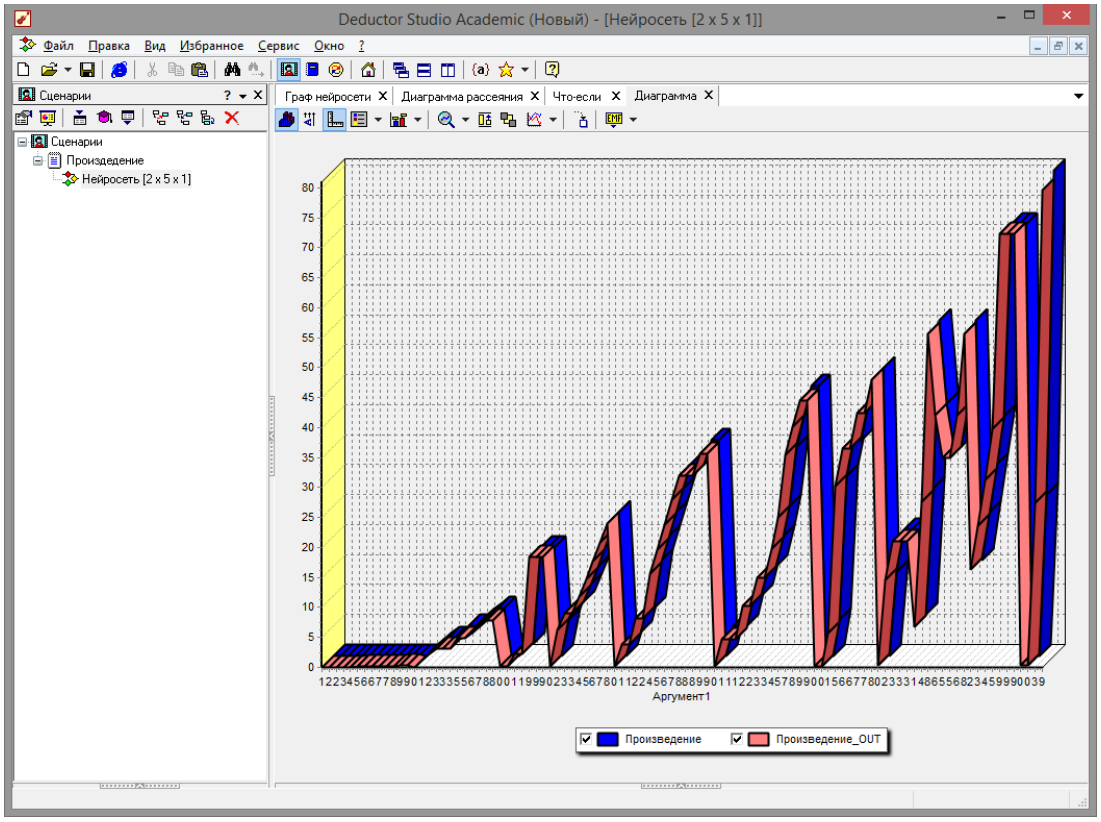


Рис. 3.10 - Сравнение эталонных данных с прогнозируемыми

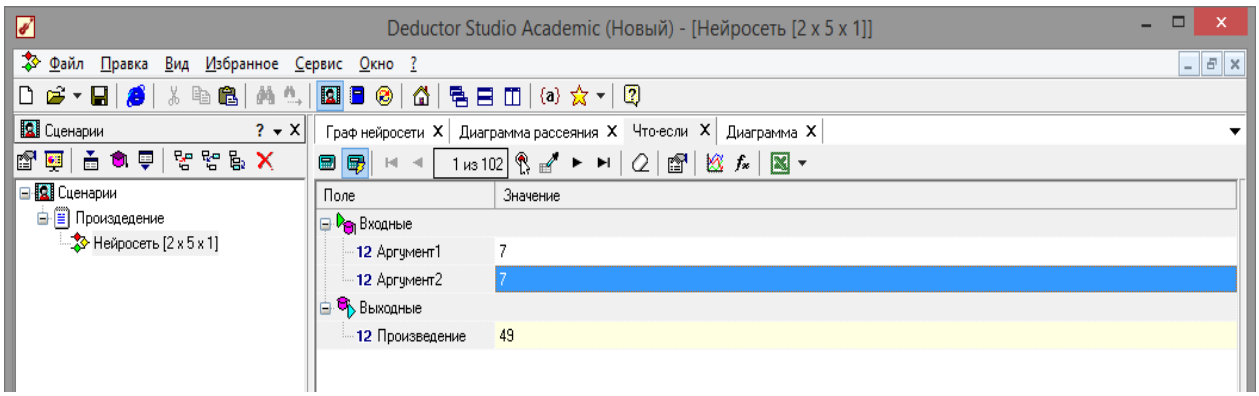


Рис. 3.11 - Инструмент «Что-Если»

Далее введем в окно «Что-Если» аргументы, которые как следует из диаграммы и диаграммы рассеивания большую ошибку. Как видим по произведению, данные диаграммы верни относительно ошибки.

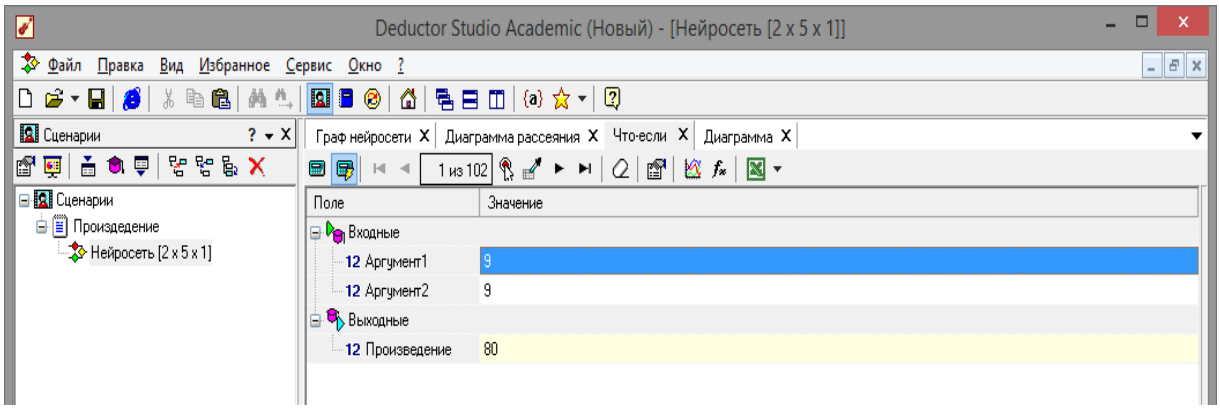


Рис. 3.12 - Ошибки прогнозирования

Вид построенной сети можно посмотреть, выбрав визуализатор «Граф нейронной сети» (рис. 3.13).

Данный пример показал, как можно построить модель прогноза, используя нейронную сеть. Пример показал, что для построения нет необходимости в строгой математической спецификации модели, что особенно ценно при анализе плохо формализуемых процессов. А большинство бизнес задач плохо формализуется. Это означает, что наличие достаточно развитых и удобных инструментальных программных средств позволяет аналитику при построении модели прогнозируемого процесса руководствоваться такими понятиями, как опыт и интуиция.

Настройки мастера позволяют увидеть широкие возможности *Deductor Studio* касательно структуры сети, способов обучения и т.д. Аналитику предоставляется широкие возможности по настройке

нормализации столбцов, разбиения данных на обучающее и тестовое множество, определения структуры сети, количества слоев и нейронов в каждом слое, выборе функции активации и ее параметров, выборе различных алгоритмов обучения и настройки их параметров.

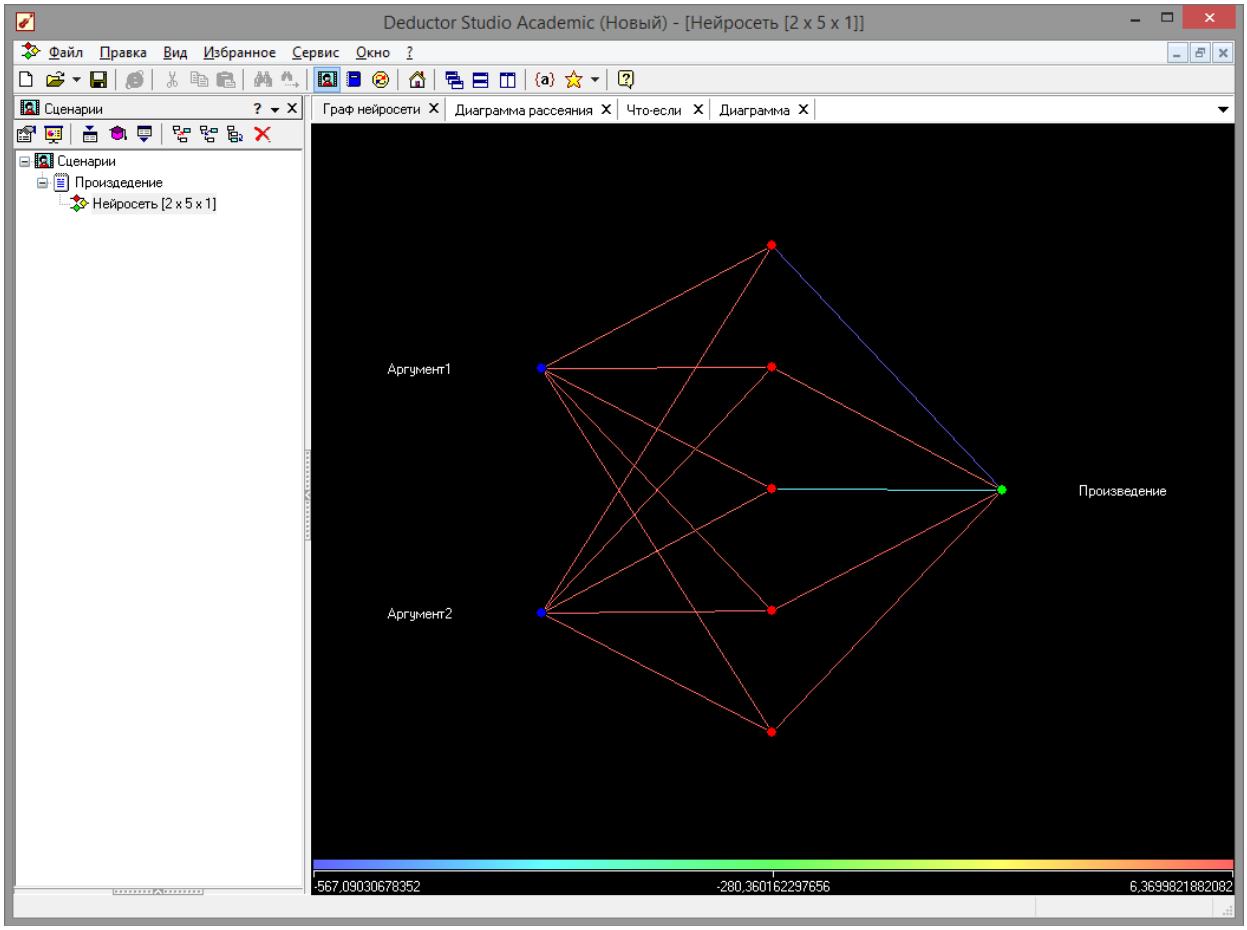


Рис. 3.13 - Граф нейронной сети

Все это позволяет построить модель, описывающую практически любые закономерности. Также было показано, как можно спрогнозировать результат, введя любые значения входных факторов, используя визуализатор «Что-если». Качество подготовки данных для модели, а также качество самой модели аналитик может оценить разными способами: посмотреть диаграмму рассеяния, провести ряд экспериментов при помощи «Что-если», построить гистограмму распределения ошибки и т.п.

3.2 Прогнозирование данных на основе временного ряда

Прогнозирование результата на определенное время вперед, основываясь на данных за прошедшее время – задача, встречающаяся довольно часто (к примеру, перед большинством торговых фирм стоит задача оптимизации складских запасов, для решения которой требуется знать, чего и сколько должно быть продано через неделю, и т.п.; задача предсказания стоимости акций какого-нибудь предприятия через день и т.д. и другие подобные вопросы). *Deductor Studio* предлагает для этого инструмент «Прогнозирование».

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.

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. Исходные данные по продажам находятся в файле «Продажи.txt» Выполним импорт данных из файла (рис. 3.14).

После импорта данных воспользуемся диаграммой для их просмотра. На ней видно, что данные содержат аномалии (выбросы) и шумы, за которыми трудно разглядеть тенденцию. Поэтому перед прогнозированием необходимо удалить аномалии и сгладить данные. Сделать это можно при помощи спектральной обработки. Запустим мастер обработки (рис. 3.15), выберем в качестве обработки данных спектральную обработку и перейдем на следующий шаг мастера. Следующий шаг отвечает за удаление аномалий из исходного набора. Выберем поле для обработки «КОЛИЧЕСТВО» и укажем для него вычитание шума (степень вычитания – малая).

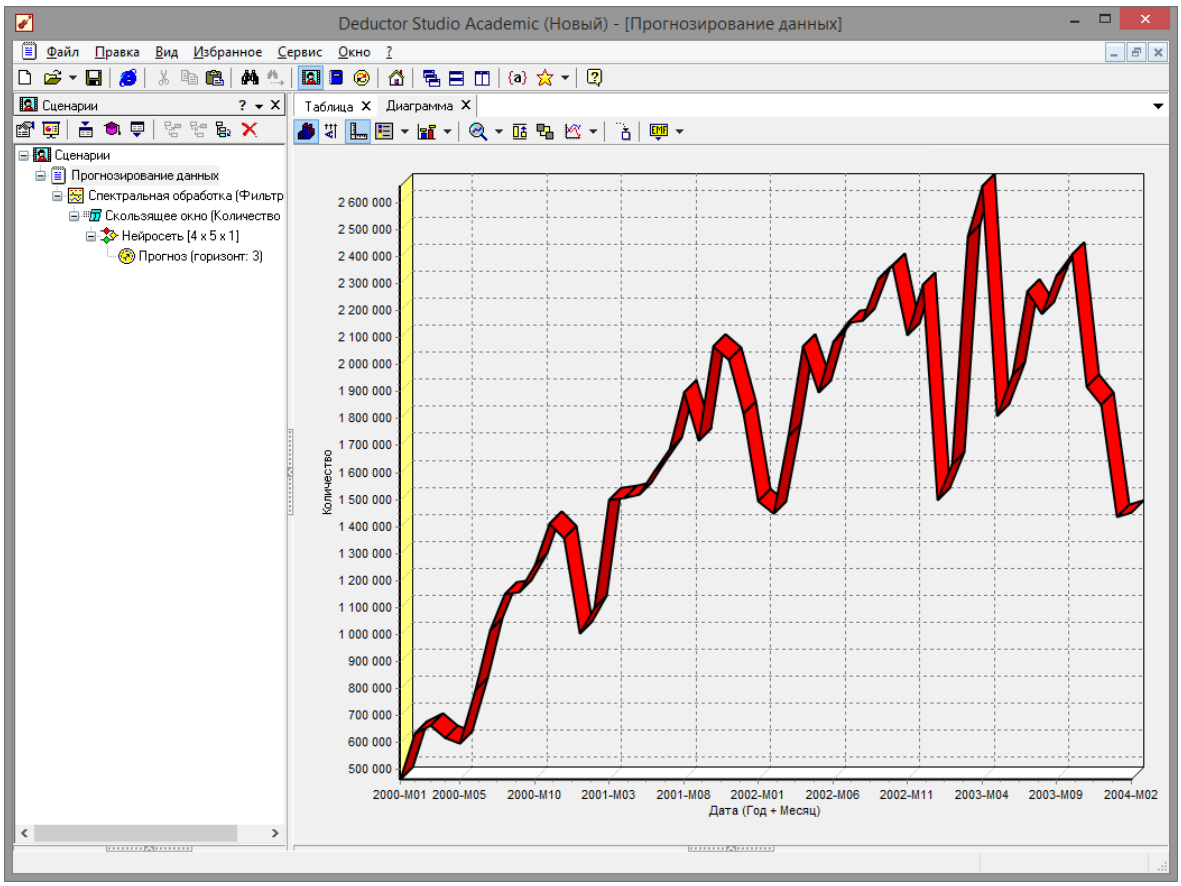


Рис. 3.14 - Диаграмма продаж

На следующем шаге запустим обработку, нажав на «пуск» и посмотрим полученный результат (рис. 3.16). Видно, что данные сгладились, аномалии и шумы исчезли. Также видна тенденция. Теперь перед аналитиком встает вопрос, а как, собственно, прогнозировать временной ряд. Во всех предыдущих примерах мы сталкивались с ситуацией, когда есть входные столбцы - факторы и есть выходные столбцы – результат. В данном случае столбец один.

Строить прогноз на будущее необходимо, основываясь на данных прошлых периодов. Предполагается, что количество продаж на следующий месяц зависит от количества продаж за предыдущие месяцы. Входными факторами для модели могут быть продажи за текущий месяц, продажи за месяц ранее и т.д., а результатом должны быть продажи за следующий месяц. Здесь явно необходимо трансформировать данные к скользящему окну.

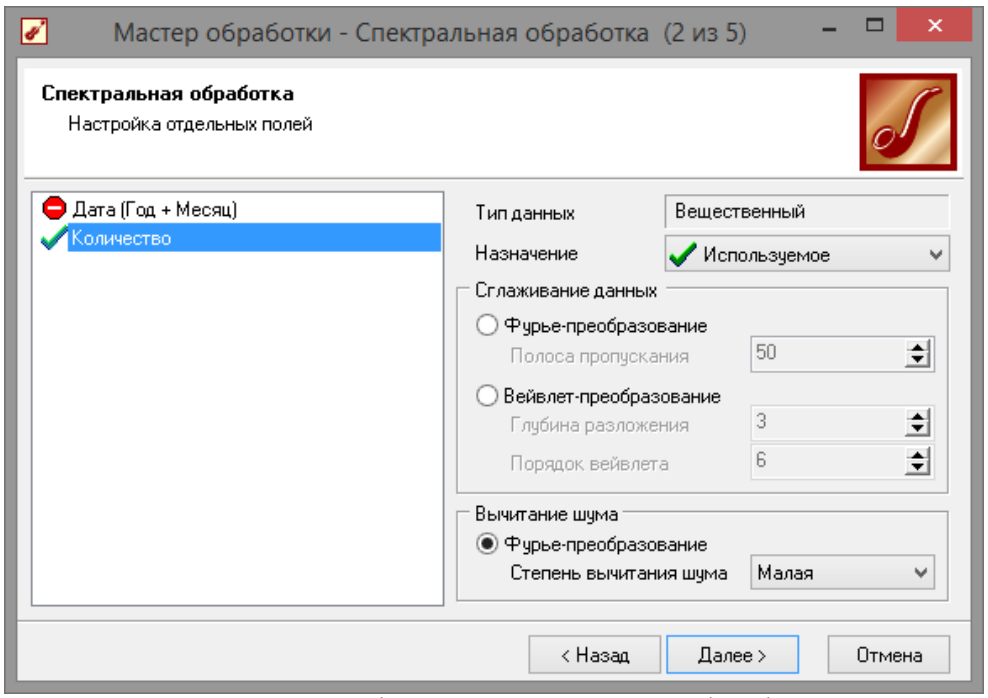


Рис. 3.15 - Выбор спектральной обработки

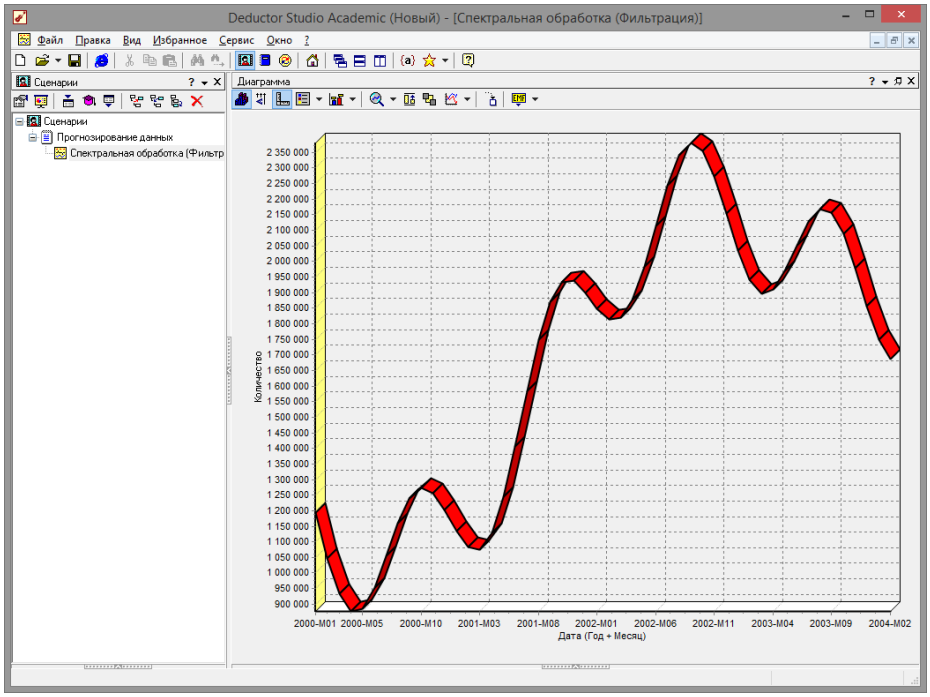


Рис. 3.16 - Результат спектральной обработки

Запустим мастер обработки, выберем в качестве обработчика скользящее окно и перейдем на следующий шаг. Было решено строить прогноз на неделю вперед, основываясь на данных за 12, 11 месяцев назад, два месяца назад и месяц 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 (рис. 3.17).

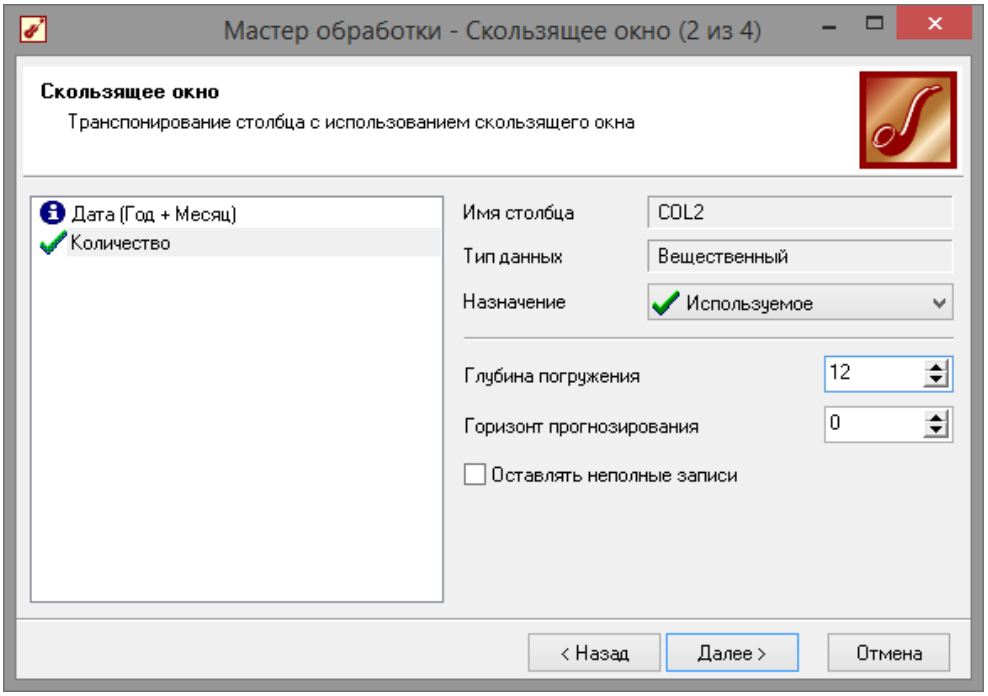


Рис. 3.17 - Скользящее окно

Просмотреть полученные данные можно в виде таблицы. 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 буден указан столбец «КОЛИЧЕСТВО».

Перейдем непосредственно к самому построению модели прогноза. Откроем мастер обработки и выберем в нем нейронную сеть (рис. 3.19). На втором шаге мастера, согласно с принятым ранее решениям, установим в качестве входных поля «КОЛИЧЕСТВО-12»,

«КОЛИЧЕСТВО-11», «КОЛИЧЕСТВО-2» и «КОЛИЧЕСТВО-1», а в

качестве выходного - «КОЛИЧЕСТВО». Остальные поля сделаем информационными.

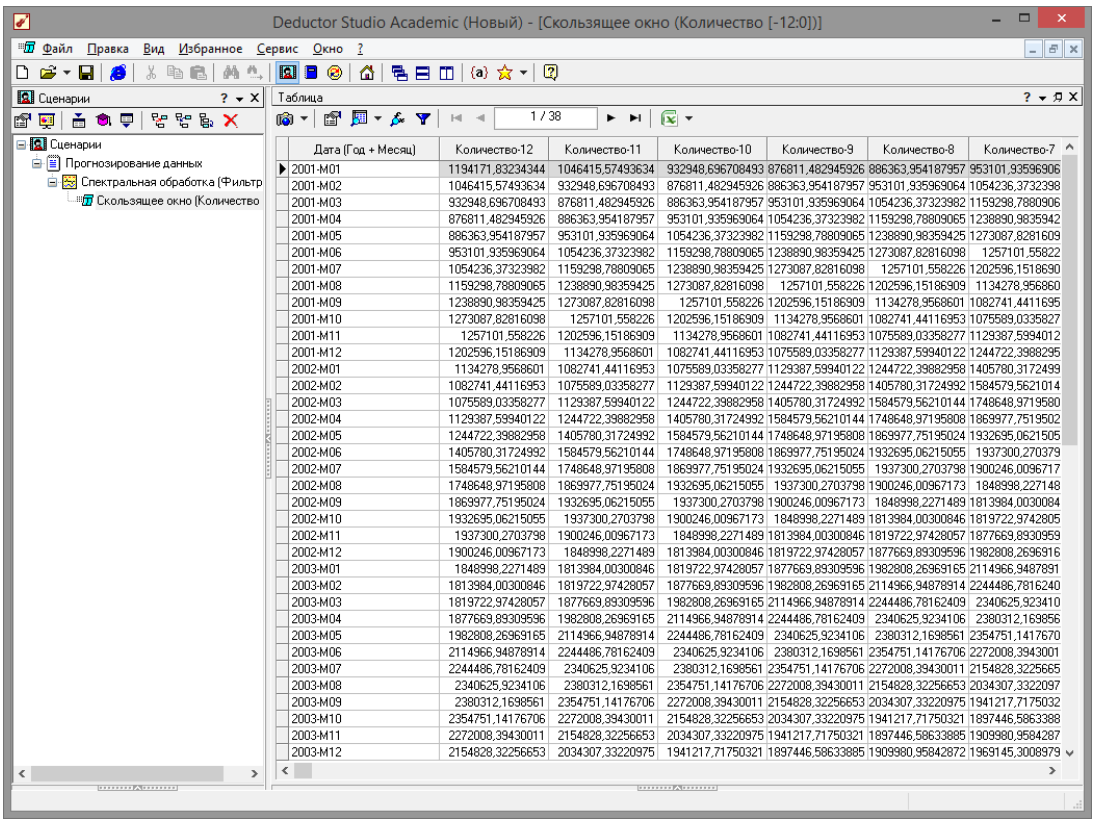


Рис. 3.18 - Таблица скользящего окна

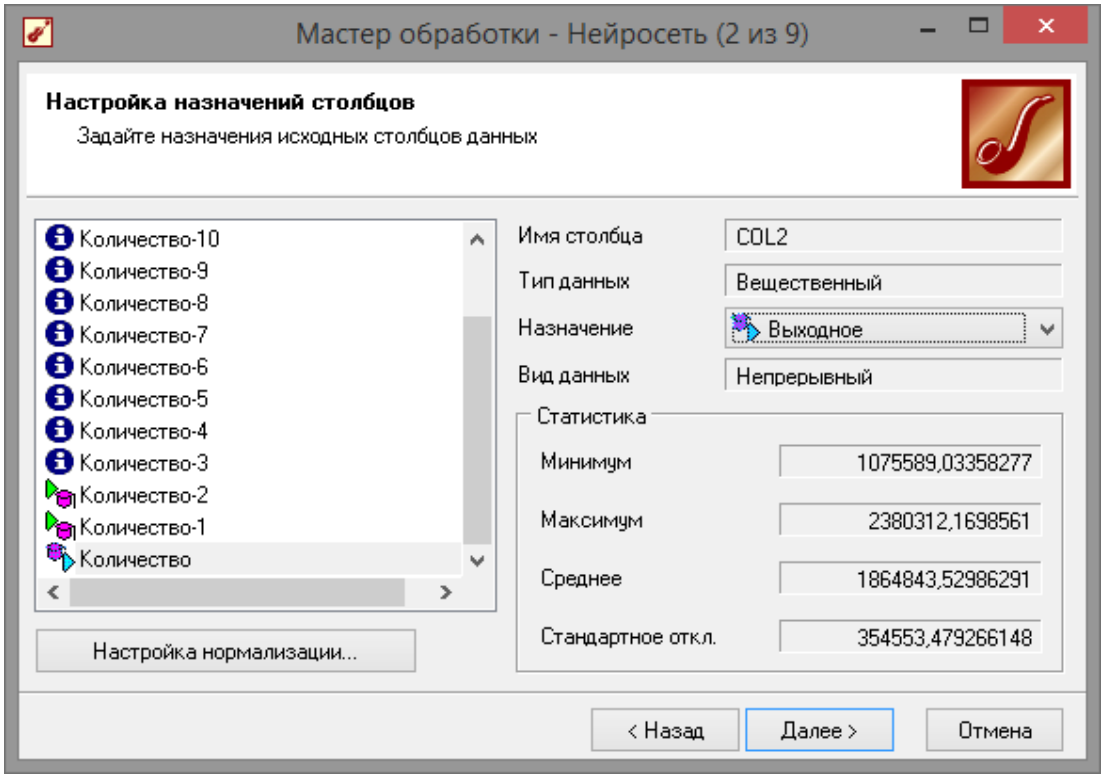


Рис. 3.19 - Настройки мастера спектральной обработки

Оставив остальные параметры построения модели по умолчанию, только количество нейронов скрытого слоя поставим равным 5

(рис. 3.20), обучим нейросеть (см. пример «прогнозирование

умножения с помощью нейронной сети») (рис. 3.21). После построения модели для просмотра качества обучения представим полученные данные в виде диаграммы и диаграммы рассеяния. В мастере настройки диаграммы выберем для отображения поля

«КОЛИЧЕСТВО» и «КОЛИЧЕСТВО_OUT» - реальное и спрогнозированное значение. Результатом будет два графика, показанные на рис.3.22.

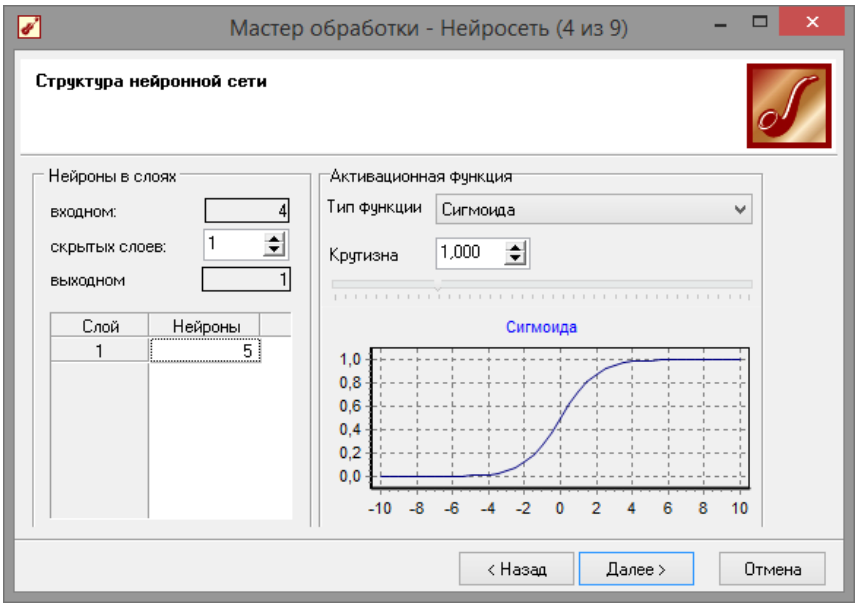


Рис. 3.20 - Настройки нейросети

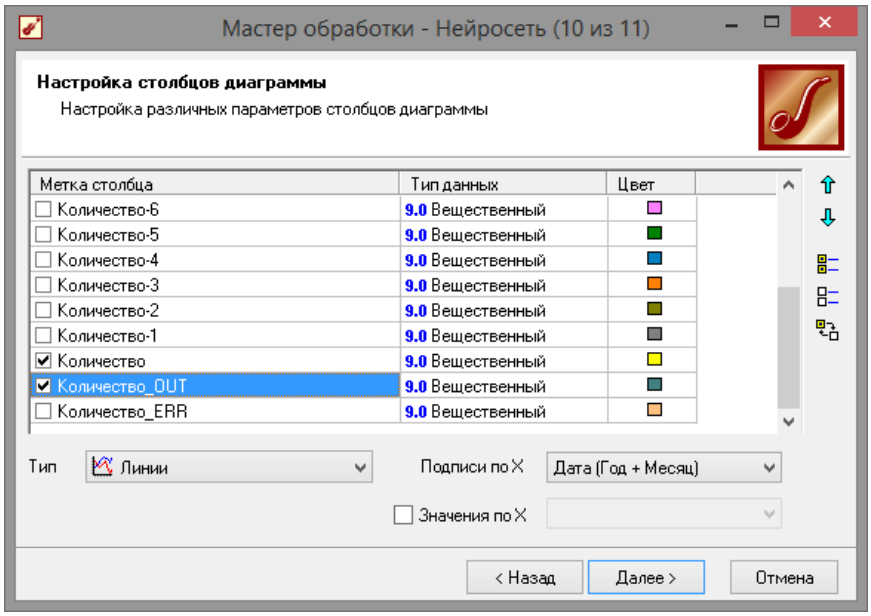


Рис. 3.21 - Обучение нейронной сети

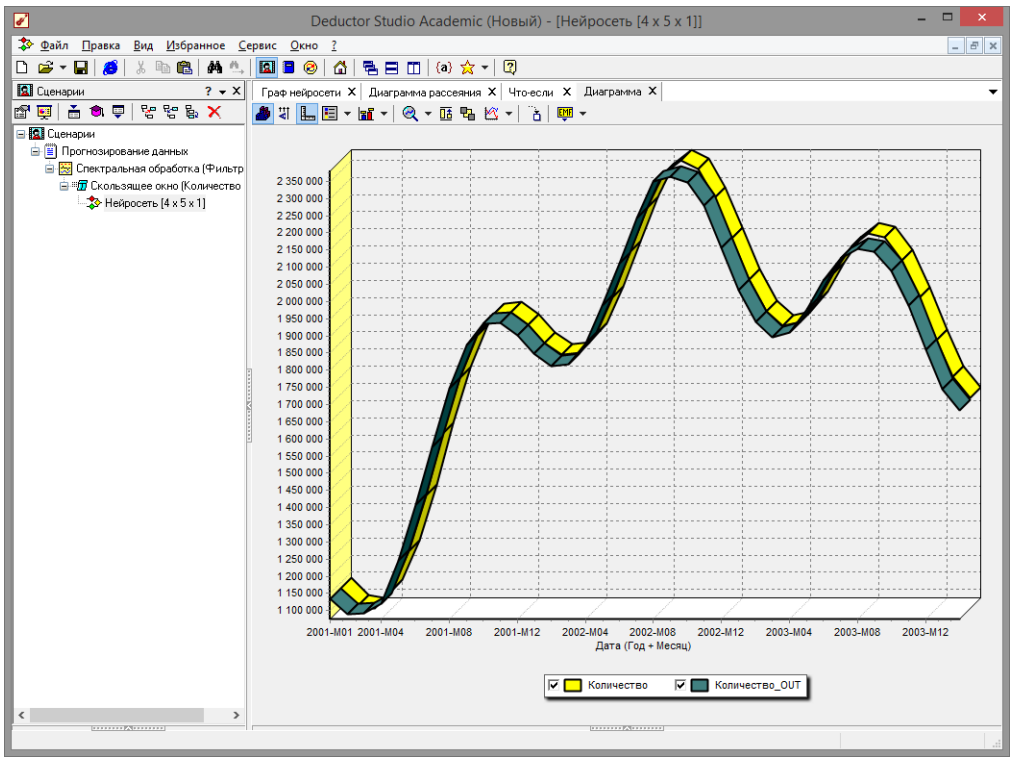


Рис. 3.22 - Сравнение эталонных данных с прогнозом

Диаграмма рассеяния более наглядно показывает качество обучения (рис. 3.23).

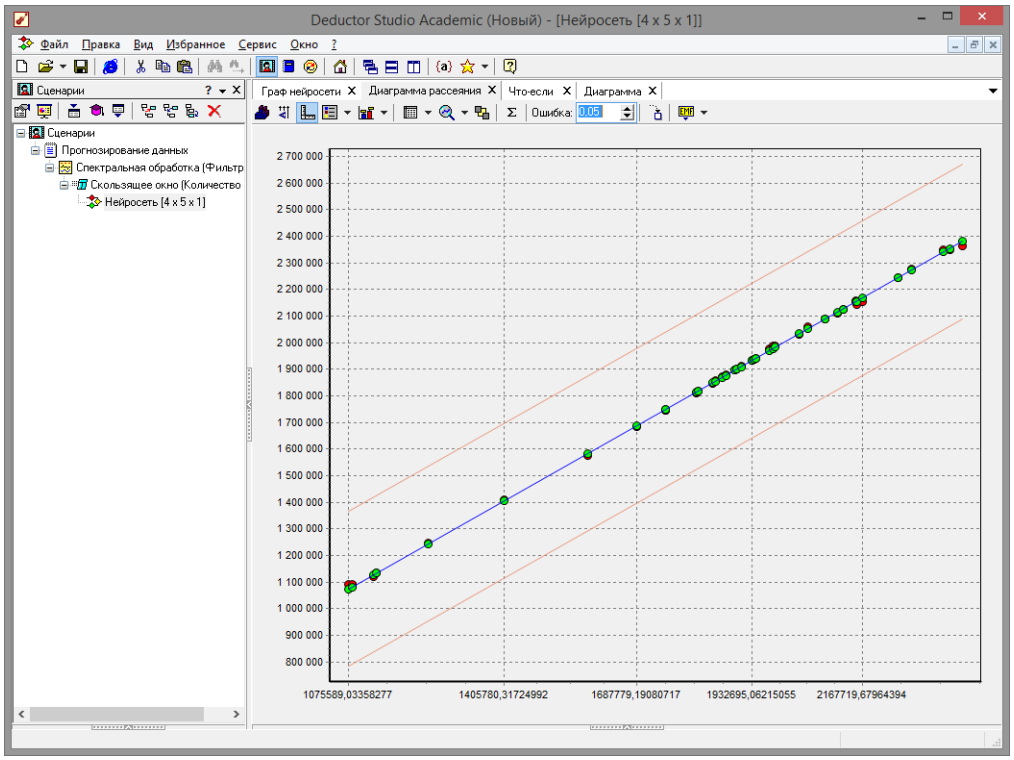


Рис. 3.23 - Диаграмма рассеяния

Нейросеть обучена, теперь осталось самое главное – построить требуемый прогноз. Для этого открываем мастер обработки (рис. 3.24) 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 (рис. 3.25). Мастер сам верно настроил все переходы, поэтому остается только указать горизонт прогноза (на сколько вперед будем прогнозировать) равным трем, а также, для наглядности, необходимо добавить к прогнозу исходные данные, установив в мастере соответствующий флажок.

После этого необходимо в качестве визуализатора выбрать диаграмму прогноза, которая появляется только после прогнозирования временного ряда. В мастере настройки столбцов диаграммы прогноза необходимо указать в качестве отображаемого столбец «КОЛИЧЕСТВО». Теперь аналитик может дать ответ на вопрос, какое количество товаров будет продано в следующем месяце и даже два месяца спустя (рис 3.26). Масштабировав результат и включив метки, можно увидеть расчетные значение на 3 месяца вперед (рис.3.27).

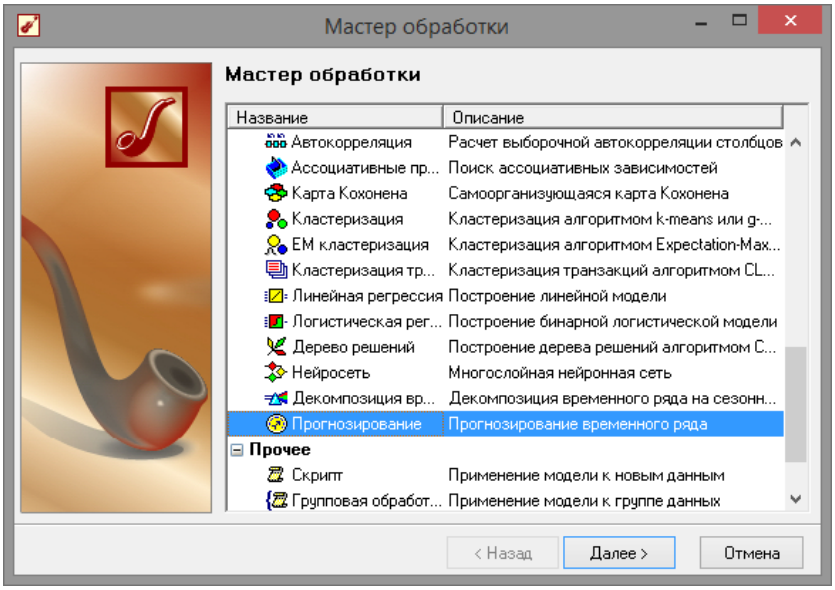


Рис. 3.24 - Мастер прогнозирования

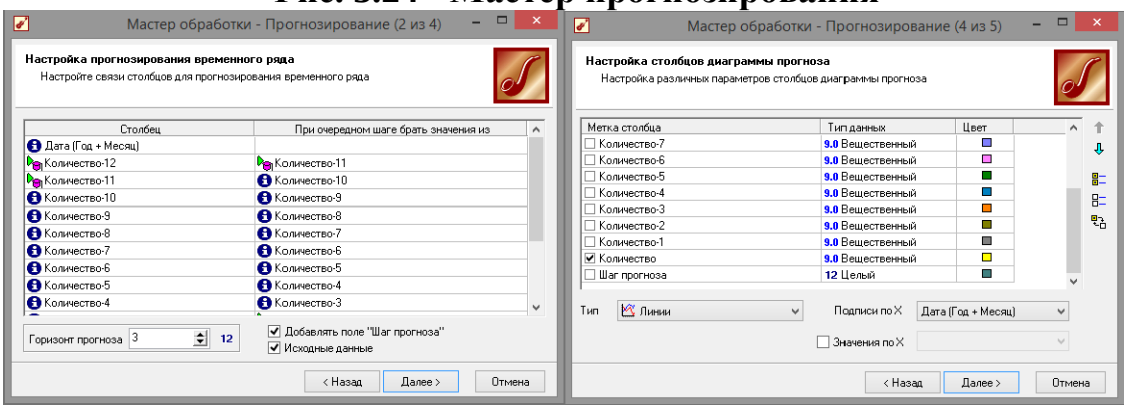


Рис. 3.25 - Мастер прогнозирования

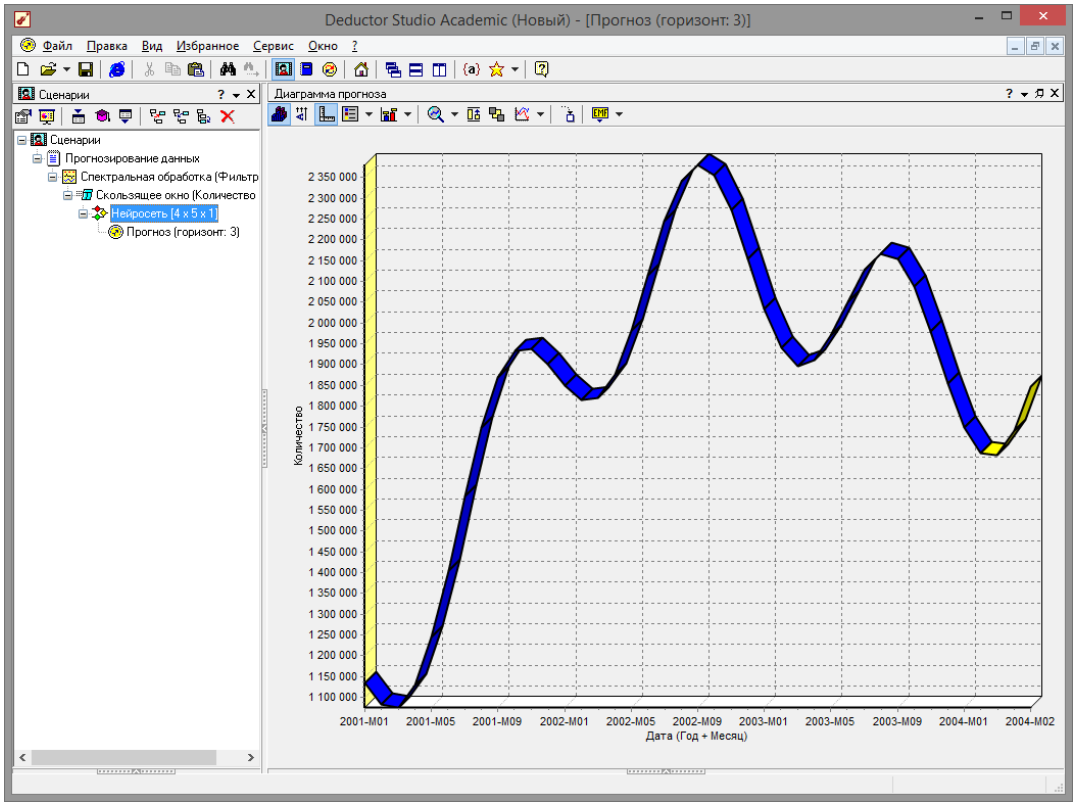


Рис. 3.26 - Расчетные значение на 3 месяца вперед

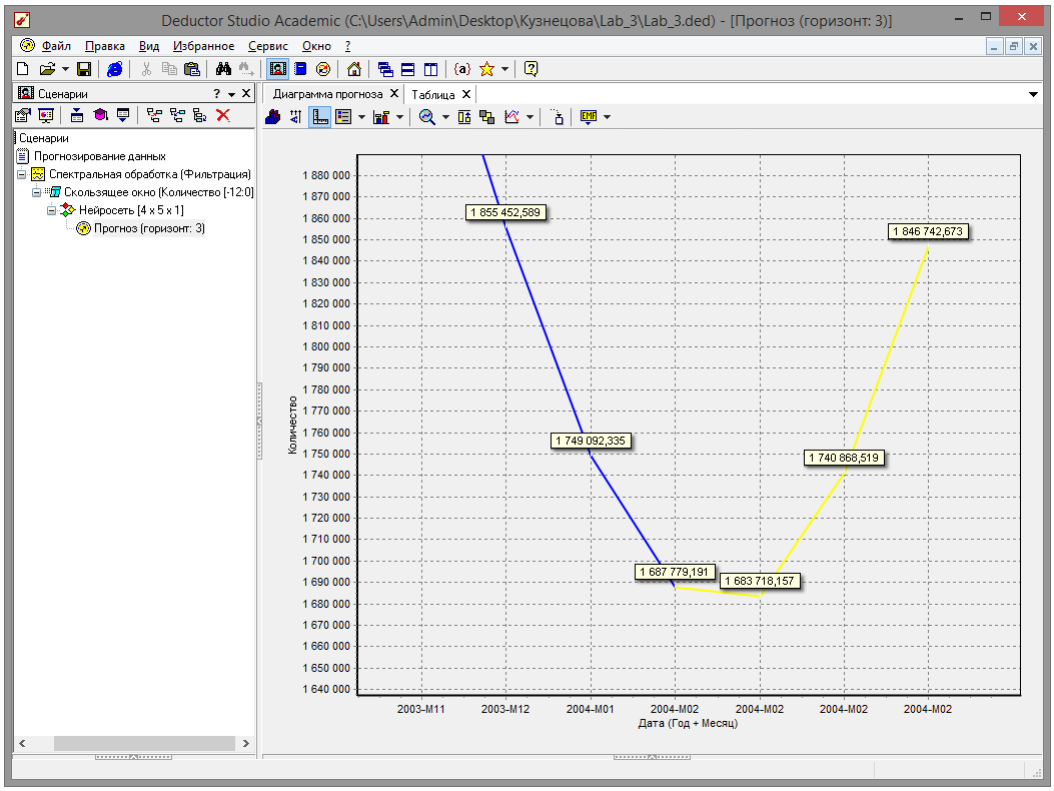


Рис. 3.27. Результаты мастера прогнозирования

После завершения анализа данные можно экспортировать. Так как

данная версия является бесплатной для образовательных целей, то данные можно выгрузить либо в текстовый файл, либо в собственный проект программы с расширением «*.*ded*». Мастер экспорта имеет точно такие же настройки, как и мастер импорта. Более того, если экспорт данных совершить в текстовый фал, то далее данные можно скопировать в файл табличного процессора *Excel,* и достаточно комфортно с ними работать.

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, а также экспорт результатов во внешний файл.

3.3 Задание на самостоятельную работу

Получить от преподавателя вариант задания для прогнозирования (изменение курса валют, график синусоиды, прогнозирование суммы или разности чисел и др.).

Контрольные вопросы

Что такое временной ряд?

В какие форматы можно экспортироват данные *Deductor Academic*?

Что такое обучающая и тестовая выборка?

Какие инструменты можно использовать д прогнозирования?

5. Для чего служит диаграмма рассеяния?

Лабораторная работа 4

Нейросетевые технологии в интеллектуальном анализе данных

Цель работы: изучить кластеризацию с помощью самоорганизующихся карт Кохонена в аналитическом пакете Deductor Academic.

Программа работы

1. Произвести импорт данных из подготовленного файла.
2. С помощью Карты Кохонена выполнить кластеризацию на данных контрольного примера.
3. Выполнить задачу кластеризацию для данных по индивидуальному заданию.

Методические указания по выполнению работы

4.1 Общие понятия о самоорганизующихся картах Кохонена

Самоорганизующаяся карта Кохонена (англ. *Self-organizing map - SOM*) – соревновательная нейронная сеть с обучением без учителя, выполняющая задачу визуализации и кластеризации. Идея сети предложена финским учёным Т. Кохоненом. Является методом проецирования многомерного пространства в пространство с более низкой размерностью (чаще всего, двумерное), применяется также для решения задач моделирования, прогнозирования и др.

Самоорганизующаяся карта состоит из компонентов, называемых узлами или нейронами. Их количество задаётся аналитиком. Каждый из узлов описывается двумя векторами. Первый – т. н. вектор веса *m,* имеющий такую же размерность, что и входные данные. Второй — вектор *r,* представляющий собой координаты узла на карте. Обычно

узлы располагают в вершинах регулярной решётки с квадратными или шестиугольными ячейками.

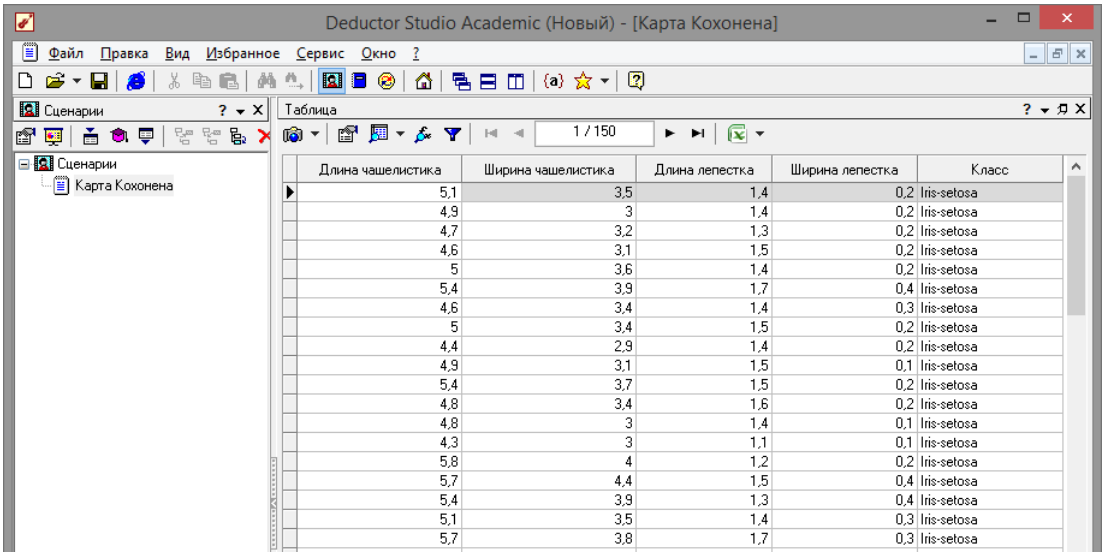
Самоорганизующаяся карта Кохонена является разновидностью нейронной сети. Она применяется, когда необходимо решить задачу кластеризации, т.е. распределить данные по нескольким кластерам. Алгоритм определяет расположение кластеров в многомерном пространстве факторов. Исходные данные будут относиться к какому-либо кластеру в зависимости от расстояния до него. Многомерное пространство трудно для представления в графическом виде. Механизм же построения карты Кохонена позволяет отобразить многомерное пространство в двумерном, которое более удобно и для визуализации, и для интерпретации результатов аналитиком. Также с помощью построенной карты Кохонена можно решить и задачу прогнозирования. В этом случае результирующее поле (то, которое необходимо спрогнозировать) в построении карты не участвует. После кластеризации, используя диаграмму «Что-если», можно провести эксперимент. Алгоритм определяет точку пространства, где расположены введенные для прогноза данные и к какому кластеру принадлежит данная точка, и подсчитывает среднее по результирующему полю всех точек этого кластера, что и будет результатом прогноза (для дискретных данных результатом прогноза является значение, больше всего встречающееся в результирующем поле всех ячеек кластера).

4.2. Пример кластеризации данных

Рассмотрим механизм кластеризации путем построения самоорганизующейся карты, основываясь на типичных характеристиках цветков. Исходная таблица находится в файле примеров «Ирисы.txt». Она содержит следующие параметры цветов:

«ДЛИНА ЧАШЕЛИСТИКА», «ШИРИНА ЧАШЕЛИСТИКА», «ДЛИНА ЛЕПЕСТКА», «ШИРИНА ЛЕПЕСТКА», «КЛАСС ЦВЕТКА». Задача состоит в том, чтобы определить по различным параметрам цветка его класс. Предполагается, что цветы одного класса имеют схожие параметры, поэтому они должны находиться в одном кластере.

Для начала необходимо импортировать данные из файла (рис. 4.1). После этого запустим, мастер обработки и выберем из списка метод обработки «Карта Кохонена» (рис. 4.2). На втором шаге мастера настроим назначения столбцов (рис. 4.3). Укажем столбцу «КЛАСС ЦВЕТКА» назначение «Выходной», а остальным – «Входной». Т.е. на основе данных о цветке будем относить его к тому или иному классу.



Длина чашелистика	Ширина чашелистика	Длина лепестка	Ширина лепестка	Класс
5,1	3,5	1,4	0,2	Iris-setosa
4,9	3	1,4	0,2	Iris-setosa
4,7	3,2	1,3	0,2	Iris-setosa
4,6	3,1	1,5	0,2	Iris-setosa
5	3,6	1,4	0,2	Iris-setosa
5,4	3,9	1,7	0,4	Iris-setosa
4,6	3,4	1,4	0,3	Iris-setosa
5	3,4	1,5	0,2	Iris-setosa
4,4	2,9	1,4	0,2	Iris-setosa
4,9	3,1	1,5	0,1	Iris-setosa
5,4	3,7	1,5	0,2	Iris-setosa
4,8	3,4	1,6	0,2	Iris-setosa
4,8	3	1,4	0,1	Iris-setosa
4,3	3	1,1	0,1	Iris-setosa
5,8	4	1,2	0,2	Iris-setosa
5,7	4,4	1,5	0,4	Iris-setosa
5,4	3,9	1,3	0,4	Iris-setosa
5,1	3,5	1,4	0,3	Iris-setosa
5,7	3,8	1,7	0,3	Iris-setosa

Рис. 4.1 - Импортированные данные

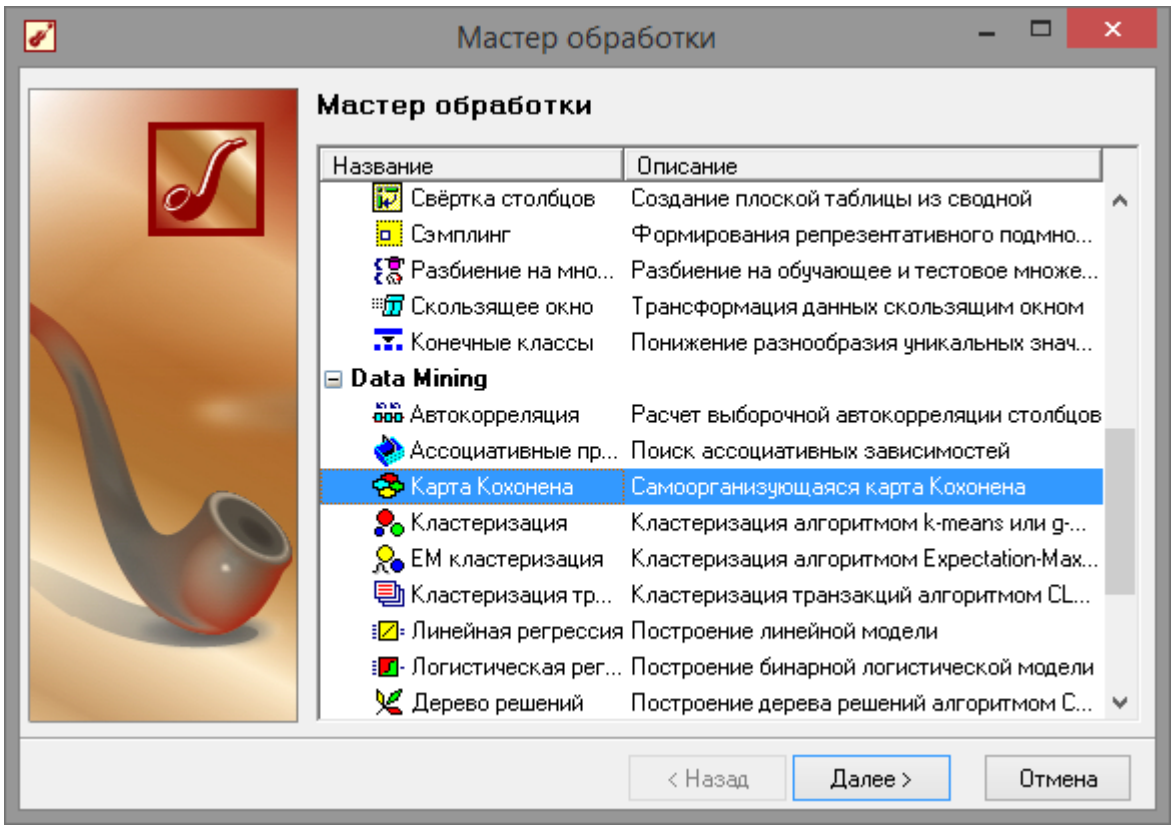


Рис. 4.2 - Мастер обработки «Карта Кохонена»

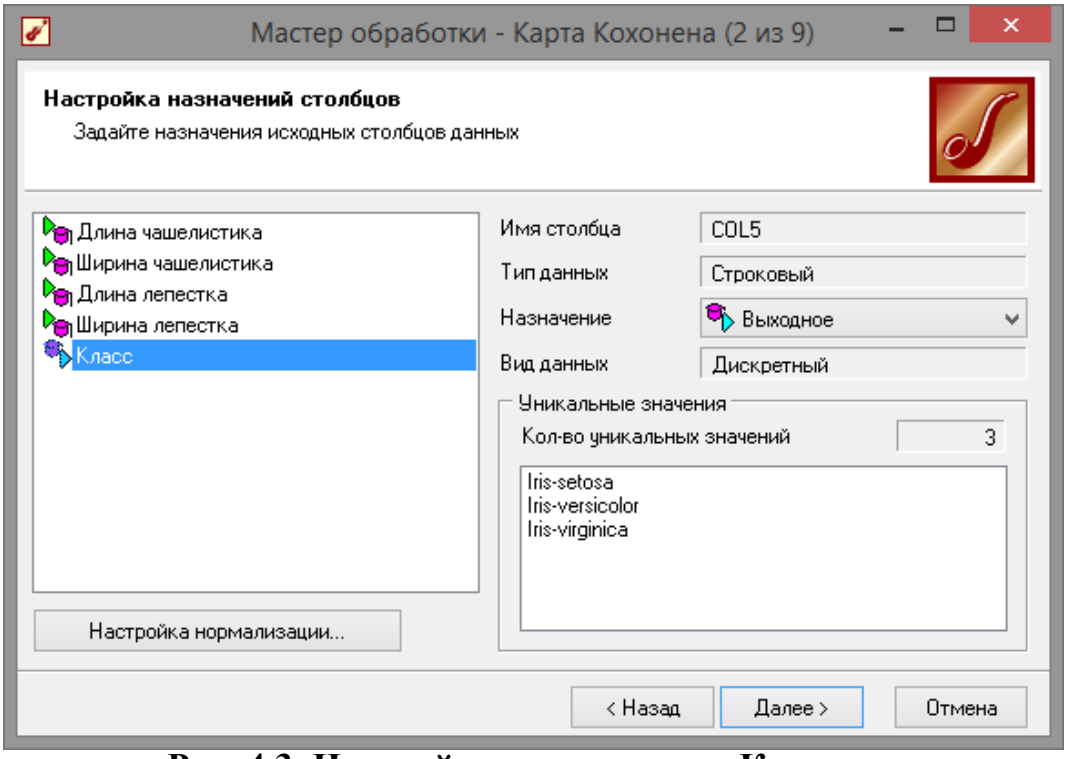


Рис. 4.3. Настройки мастера карт Кохонена

На третьем шаге мастера необходимо настроить способ разделения исходного множества данных на тестовое и обучающее, а также количество примеров в том и другом множестве. Укажем, что данные обоих множеств

берутся случайным образом, зададим размер тестового множества равным десяти примерам, путем изменения значения столбца «Размер в строках» строки «Тестовое множество» (рис. 4.4).

Следующий шаг предлагает настроить параметры карты (количество ячеек по X и по Y, их форму) и параметры обучения (способ начальной инициализации, тип функции соседства, перемешивать ли строки обучающего множества и количество эпох, через которые необходимо перемешивание). Значения по умолчанию вполне подходят (рис. 4.5).

На пятом шаге мастера необходимо настроить параметры остановки обучения. Оставим параметры по умолчанию (рис. 4.6).

На шестом шаге настраиваются остальные параметры обучения – способ начальной инициализации, тип функции соседства и также параметры кластеризации – автоматическое определение числа кластеров с соответствующим уровнем значимости либо фиксированное количество кластеров предоставляется возможность настроить интервалы обучения. Каждый интервал задается количеством эпох, радиусом обучения и скоростью обучения. Укажем фиксированное количество кластеров, равное трем (рис. 4.7).

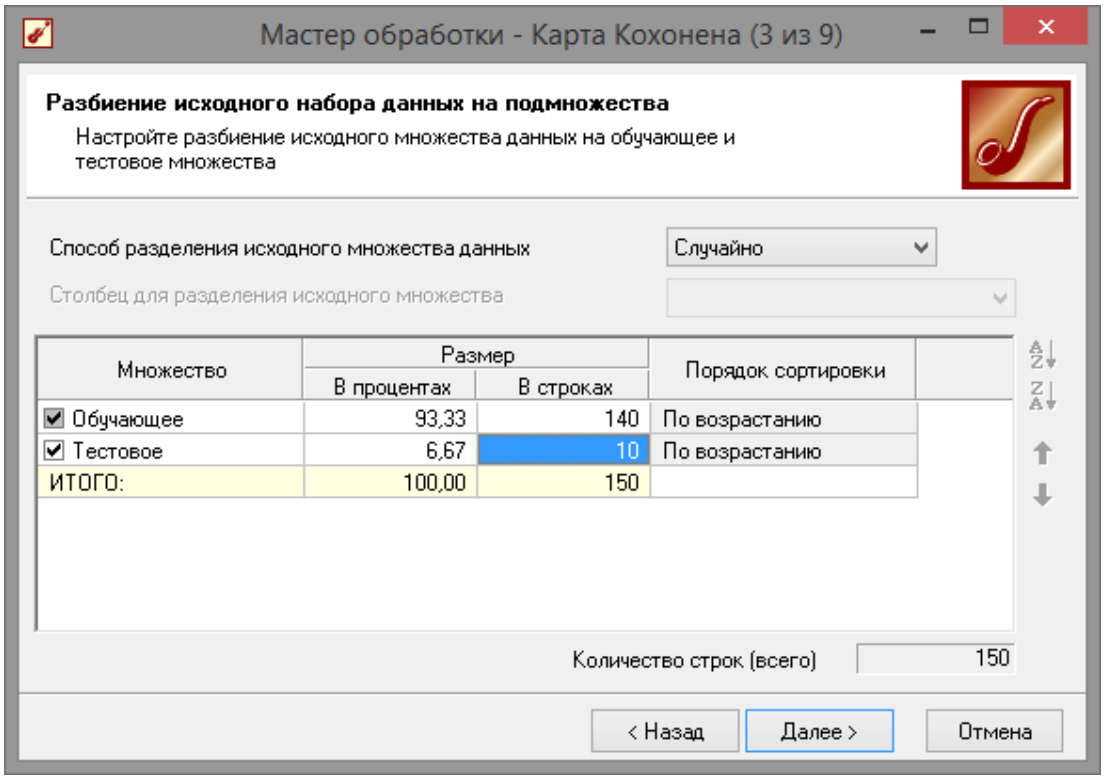


Рис.4.4 - Настройки тестового и обучающего множества

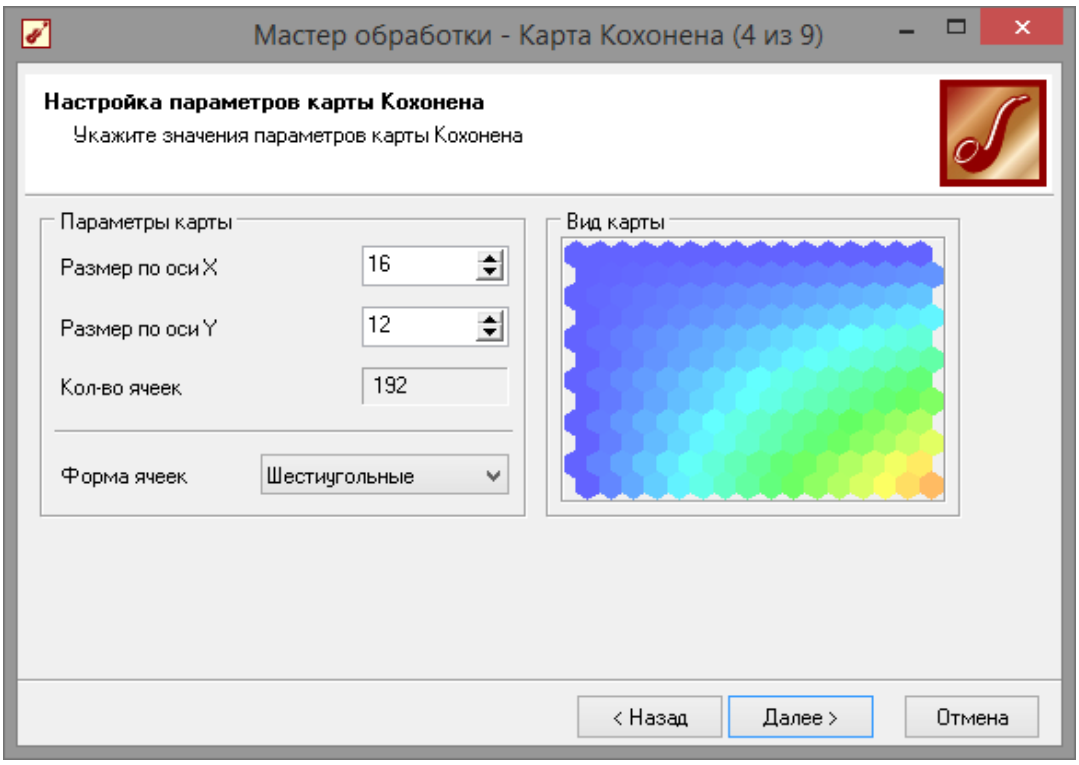


Рис. 4.5 - Настройки значения параметров карт

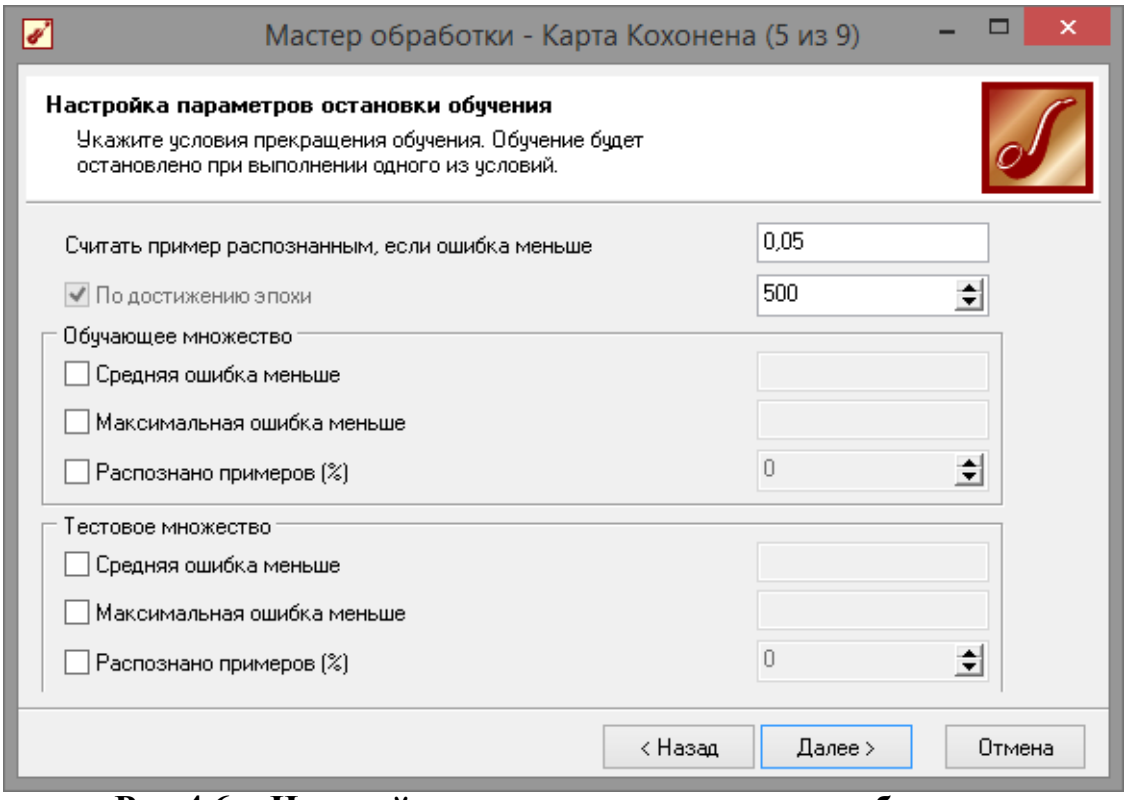


Рис.4.6 - Настройки параметров остановки обучения

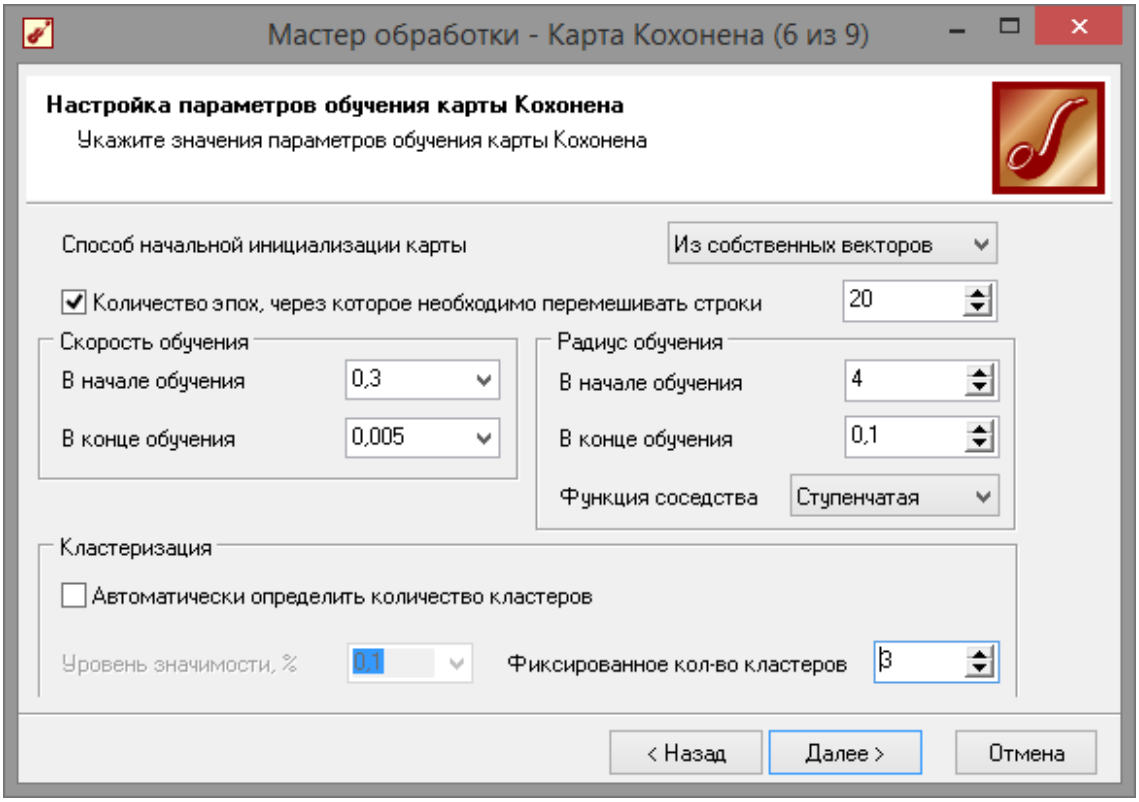


Рис. 4.7 - Настройки параметров обучения карты Кохонена

На седьмом шаге предлагается запустить сам процесс обучения (рис. 4.8). Во время обучения можно посмотреть количество распознанных примеров и текущие значения ошибок. Здесь необходимо нажать на кнопку пуск и дождаться завершения процесса обработки.

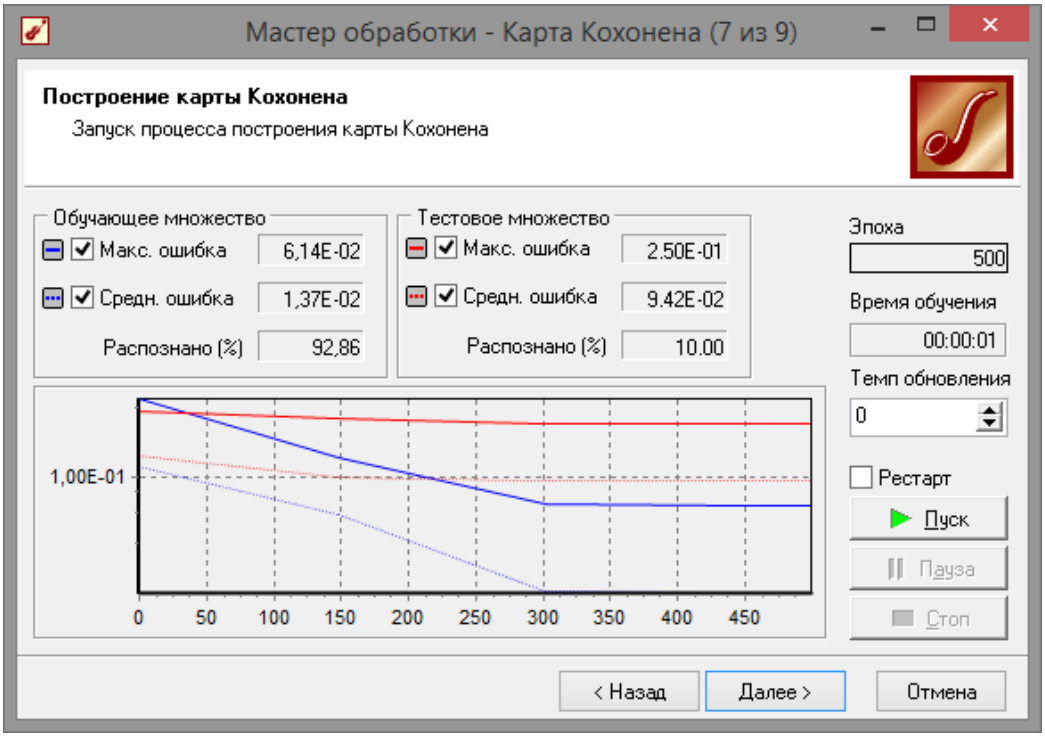


Рис. 4.8 - Процесс обучения

После этого необходимо в списке визуализаторов выбрать появившуюся теперь «Карту Кохонена» для просмотра результатов кластеризации, а также визуализатор «Что-если» для прогнозирования класса цветка (рис. 4.9).

Далее, в мастере настройки отображения карты Кохонена необходимо указать, чтобы отображались все поля, также следует установить количество кластеров равным трем и поставить флажок

«Границы кластеров» (рис. 4.10).

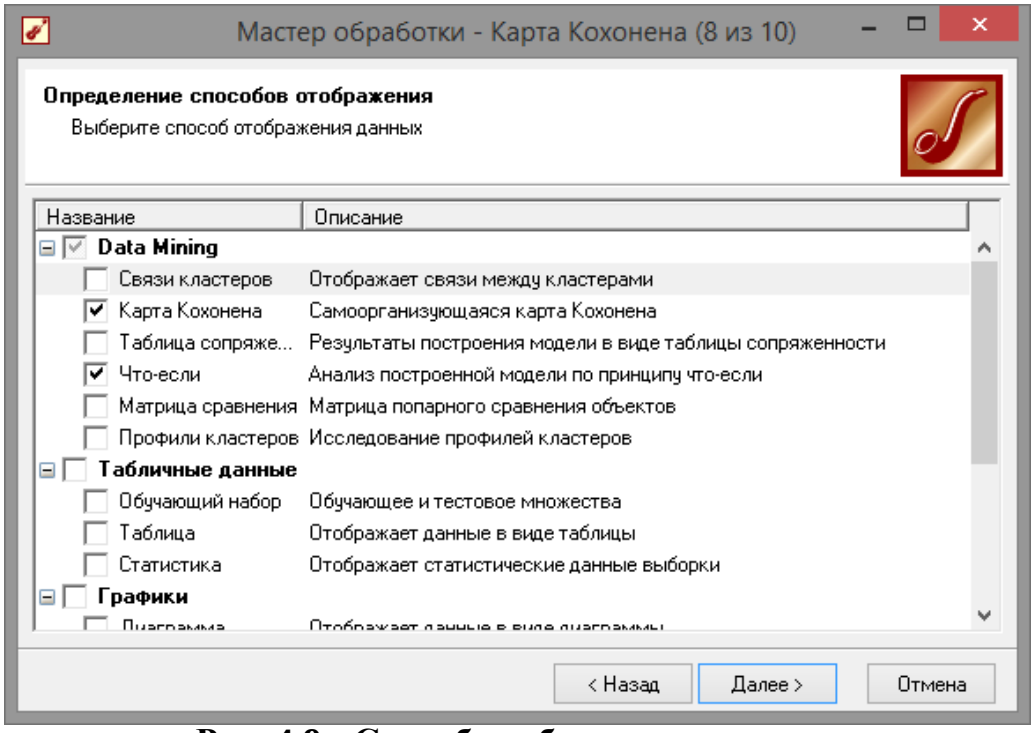


Рис. 4.9 - Способ отображения данных

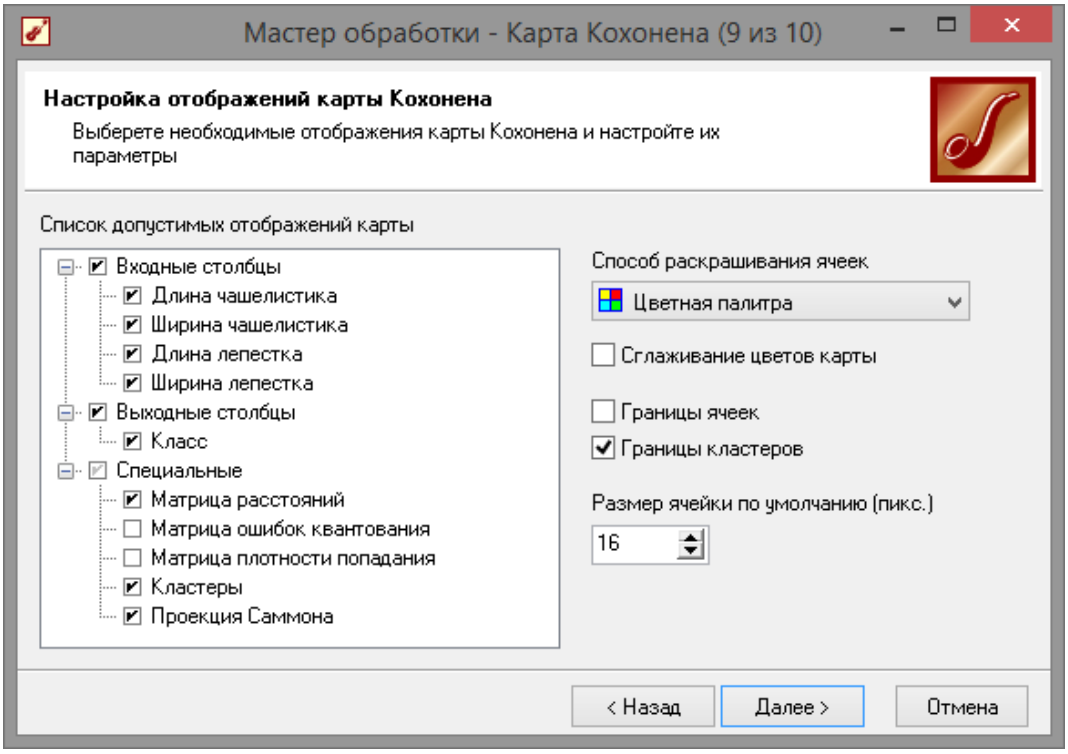


Рис. 4.10 - Настройка отображения кластеров

После этого можно увидеть полученные результаты (рис. 4.11). Качество кластеризации можно оценить, просмотрев карту «КЛАСС ЦВЕТКА». На ней видно, что большинство цветов были классифицированы правильно. Заметим, что все цветы класса *Setosa*

попали в один кластер. Это говорит о значительном отличии параметров цветов этого класса от других. Явное различие наблюдается по длине и ширине лепестка. То, что часть примеров *Virginica* попала в класс *Versicolo* и наоборот говорит о меньшем различии этих классов. На картах, в отличие от *Setosa* не видны резкие отличия параметров цветов этих двух классов. Этим как раз и объясняется «проникновение» некоторой части примеров в другой кластер.

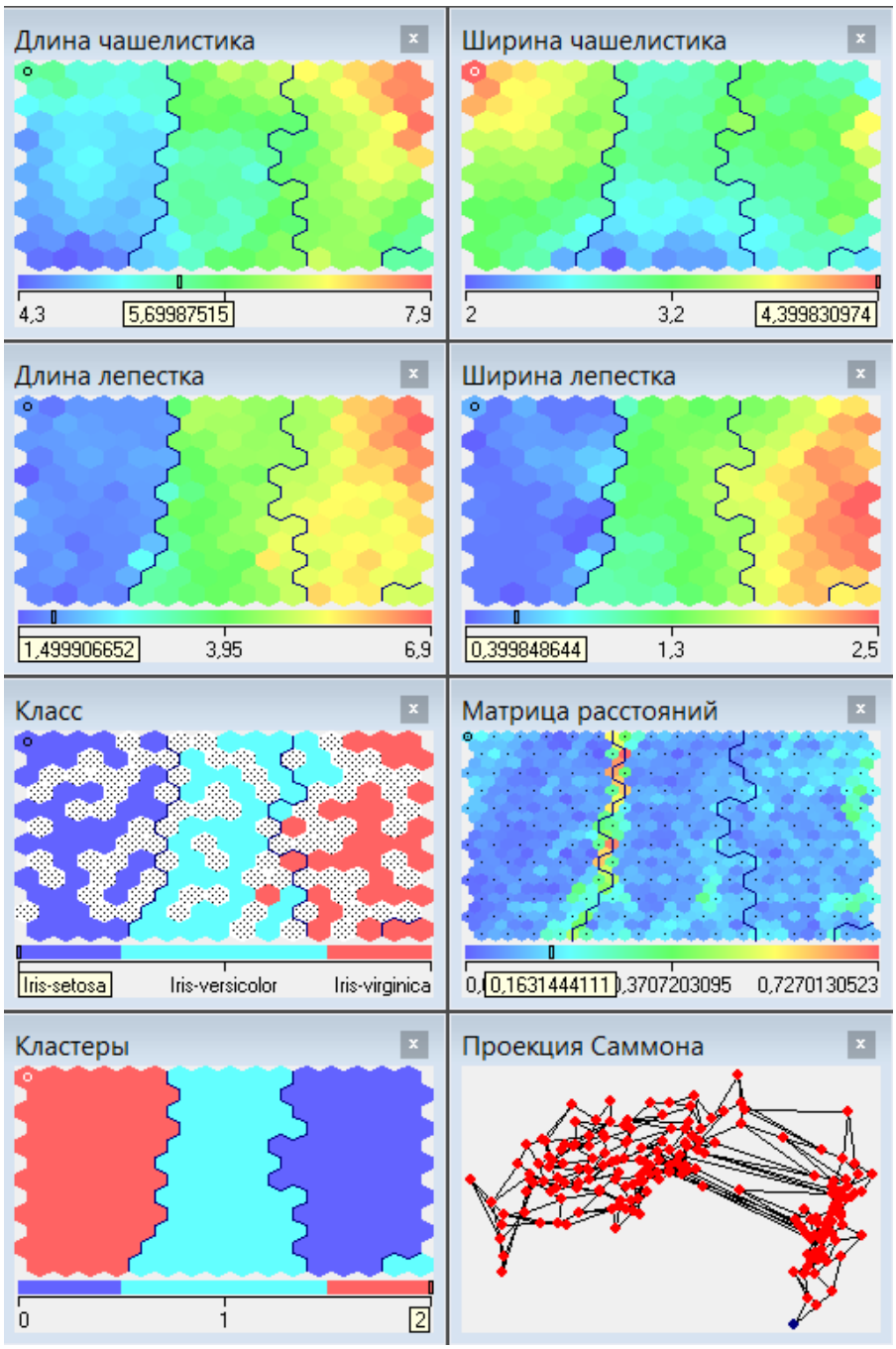


Рис. 4.11 - Карта Кохонена

Рассмотрим построенную таблицу «Что-Если». В верхней части таблицы отображаются входные поля, а в нижней – выходные и расчетные. Изменяя значения входных полей, пользователь дает команду на выполнение расчета и наблюдает за рассчитанными значениями выходов нейронной сети или дерева решений.

Расчетные поля отличаются от выходных тем, что они не существуют в исходном наборе данных и были созданы в ходе обработки. Такими полями являются, например, «Номер ячейки» или

«Номер кластера».

Каждое поле таблицы «Что-Если» представлено следующими

атрибутами:

– «Тип» – указывается значок, соответствующий типу данных поля;

– «Поле» – имя входного или выходного поля;

– «Значение» – указывается текущее значение поля.

С помощью кнопки «Показать статистику» справа от таблицы можно вывести статистику по выделенному полю. Для непрерывных полей в ней отображается следующая информация:

– «Минимум» – минимальное значение поля в выборке;

– «Максимум» – максимальное значение поля в выборке;

– «Среднее» – среднее по выборке значение поля;

– «Стандартное откл.» – среднеквадратическое отклонение значений поля по выборке.

Знание диапазона входных данных (минимума и максимума), на котором строилась модель, позволит определить область устойчивости системы. Очевидно, что, если подать на вход значения, существенно выходящие за диапазон, гарантировать правильную реакцию системы нельзя, и достоверность полученных данных может быть снижена. Если значение, присвоенное полю, выходит за границы диапазона, это поле окрашивается в красный цвет.

Для дискретных полей статистика содержит:

– «Значения» – список уникальных значений;

– «Кол-во» – число вхождений значения в выборку;

– «Итоговая информация» – общее число уникальных значений в выборке.

Для дискретных значений на вход можно подавать только значения, представленные в этом списке.

В таблице пользователь может менять лишь содержимое столбца «Значение». Это осуществляется несколькими способами:

– непосредственно ввести данные с клавиатуры;

– заполнить записями из текущей выборки (при этом вводятся записи целиком и заполняются одновременно все поля);

– выбрать значения из статистики, находящейся справа от таблицы.

Чтобы ввести значения входов с клавиатуры, нужно выбрать ячейку «Значение» для соответствующего поля, и только потом вводить данные. Чтобы войти в режим редактирования, достаточно напечатать любой символ с клавиатуры в том числе «*Enter*» либо дважды "кликнуть" мышкой по соответствующей ячейке. Дискретные значения выбираются из выпадающего списка либо путем циклического перебора в следствии двойного "клика" мышкой. Для перехода к предыдущим или последующим строкам используются клавиши со стрелками. Если

введенные вами значения выходят за диапазон значений выборки, соответствующая строка таблицы выделяется красным цветом. Если, находясь на ячейке, нажать клавишу "*Del*", то значение соответствующего входного поля будет очищено.

Для автоматического ввода в таблицу «Что-Если» записей из текущей выборки используются кнопки на панели инструментов:

– Первая запись (Ctrl+PgUp) – позволяет выбрать для загрузки в таблицу «Что-Если» первую запись выборки;

– Предыдущая запись (PgUp) – позволяет загрузить предыдущую запись;

– Загрузить запись – загружает текущую запись в соответствующие входные поля таблицы «Что-Если»;

– Загрузить из исходной выборки – выводит на экран модальное окно с таблицей, из которой можно загрузить необходимую запись;

Как видно из таблицы «Что, если» (рис. 4.12), даже данные отсутствующие в изначальной выборке определяются корректно.

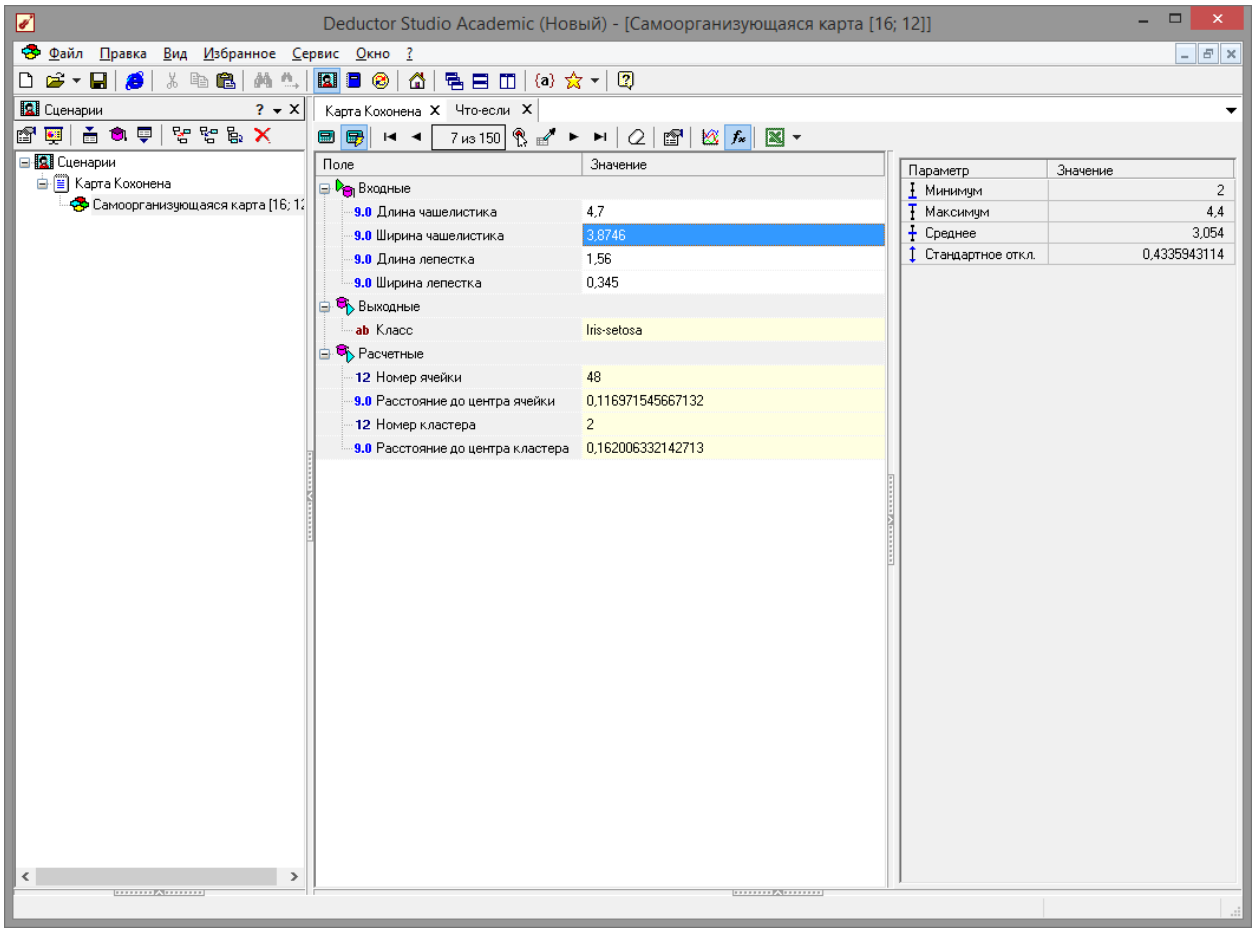


Рис. 4.12 - Таблица «Что-Если»

Данный пример показал область применения самоорганизующихся карт. Изначально имелось многомерное (четырехмерное) пространство

входных факторов. Алгоритм представил его в двумерном виде, которое удобнее анализировать. Также исходные данные были отнесены к трем кластерам, по типу цветка – «*Setosa*», «*Versicolo*», «*Virginica*». Основным визуализатором после построения является «Самоорганизующаяся карта». Здесь в первую очередь следует обратить внимание на матрицу расстояний и проекцию Саммона. На них явно видны расстояния между отдельными ячейками карты, т.е. четкие границы различных скоплений данных. Мастер предоставляет широкий набор настройки параметров обучения: настройка нормализации столбцов, настройка разбиения на тестовое и обучающее множество, настройка условий остановки обучения, настройка параметров карты и параметров обучения, настройка интервалов обучения. информация о задаче, то качество очистки данных можно увеличить на порядки.

4.3 Задание на самостоятельную работу

Получить от преподавателя вариант задания на кластеризации (классификация помидоров по диаметру плода, весу плода, количеству плодов на кусте, высоте куста, или задача классификации призывников по категориям на основе их параметров и др.).

Решить задачу при помощи карт Кохонена.

Контрольные вопросы

1. Что такое карта Кохонена?
2. Что решают задачи кластеризации?
3. Для чего служит таблица «Что-Если»?

Контрольные вопросы промежуточной аттестации (по итогам изучения курса)

1. Данные и модели их представления.
2. Системы поддержки принятия решений (СППР).
3. Роль и место интеллектуального анализа данных в СППР.
4. Задачи ИАД.
5. Алгебра матриц.
6. Функции многих переменных.
7. Необходимые и достаточные условия существования экстремумов применительно к квадратичным формам.
8. Типы шкал.
9. Допустимые преобразования в шкалах.
10. Проверка истинности утверждений.

11. Статистическая выборка.
12. Числовые характеристики распределений.
13. Комплексные числа и их применение при визуализации многомерных данных.
14. Методы и алгоритмы оцифровки графиков
15. Методы и алгоритмы обработки изображений
16. Простые и сложные признаки и способы щценки информативности
17. Алгоритмы поиска систем информативных признаков.
18. Матрица объект-признак и её статистические характеристики.
19. Проблема сжатия данных
20. Разнотипные данные и методы их обработки
21. Задача поиска логических закономерностей
22. Методы классификации и прогнозирования
23. Задачи кластерного анализа
24. Иерархические и итеративные методы кластеризации
25. Особенности кластеризации в качественных количественных шкалах
26. Кластеризация данных по матрице объект-признак.
27. Кластеризация данных по матрице матрице связи.
28. Назначение компонентного и факторного анализа.
29. Сходство и различие компонентного и факторного анализа.
30. Применение компонентного и факторного анализа к задачам ИАД.
31. Методы распознавания образов с учителем и без учителя.
32. Задачи принятия решений.
33. Метод анализа иерархий.
34. Модификации метода анализа иерархий в интересах реализации интеллектуальных подсказок пользователям.
35. Основные понятия когнитивного моделирования
36. Инструментальные средства ИАД применительно задачам СППР
37. Направления развития ИАД
39. Краткая история нейрокомпьютинга.
40. Задачи ИАД на основе искусственных нейронных сетей.
41. Место нейронных сетей среди других методов решения задач
42. Информационный подход к моделированию нейрона.
43. Биологический подход к моделированию нейрона.
44. Структура искусственной нейронные сети.
45. Структура двухкровневого персептрона, многоуровневого персептрона (МСП).
46. Особенности структуры нейронных сетей и ее влияние на свойства сети.

47. Алгоритм решения задач с помощью МСП.

48. Классификация задач, решаемых с помощью МСП.

49. Постановка задач распознавания, аппроксимации, прогнозирования. Примеры задач.

50. Топологии нейронных сетей.

МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ
Федеральное государственное автономное образовательное учреждение высшего образования
«СЕВЕРО-КАВКАЗСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ»
Невинномысский технологический институт (филиал)

Методические указания для лабораторных работ по дисциплине «Интеллектуальный анализ данных и машинное обучение»

(ЭЛЕКТРОННЫЙ ДОКУМЕНТ)

Направление подготовки 09.03.02 Информационные системы и технологии
Квалификация выпускника Бакалавр

Невинномысск 2022

Методические указания предназначены для проведения лабораторных работ по дисциплине «Интеллектуальный анализ данных и машинное обучение» для студентов направления подготовки 09.03.02 Информационные системы и технологии и соответствуют требованиям ФГОС ВПО направления подготовки бакалавров.

Составитель: доцент кафедры ИСЭА Э.Е. Тихонов

Содержание

Введение

Методические указания предназначены для подготовки и выполнения обучающимися по направлению 09.03.02 Информационные системы и технологии и самостоятельной работы, предусмотренных учебным планом по дисциплине Интеллектуальный анализ данных и машинное обучение.

Методические указания содержат описание лабораторных заданий и самостоятельной работы, задания на самостоятельную работу и правила оформления ее результатов.

Лабораторная работа — планируемая учебная, учебно-исследовательская работа обучающихся, выполняемая на аудиторных занятиях по заданию и при управлении преподавателем, при непосредственном его участии.

Самостоятельная работа — планируемая учебная, учебно-исследовательская работа обучающихся, выполняемая вне занятий по заданию и при управлении преподавателем, но без его непосредственного участия.

Практическая работа проводится с целью закрепления теоретических знаний, полученных на лекциях, применения их для выполнения практических заданий и получения практического опыта путем выполнения заданий в области интеллектуального анализа данных.

Самостоятельная работа проводится с целью:

- систематизации и закрепления полученных теоретических знаний и практических умений обучающихся;
- углубления и расширения теоретических знаний;
- формирования умений использовать нормативную, правовую, справочную документацию и специальную литературу;
- развития познавательных способностей и активности обучающихся: творческой инициативы, самостоятельности, ответственности, организованности;
- формирование самостоятельности мышления, способностей к саморазвитию, совершенствованию и самоорганизации;
- формирования общих и профессиональных компетенций
- развитию исследовательских умений.

Целью преподавания дисциплины является формирование представления о типах задач, возникающих в области интеллектуального анализа данных (Data Mining) и методах их решения, которые помогут обучающимся выявлять, формализовать и успешно решать практические задачи анализа данных, возникающие в процессе их профессиональной деятельности.

Дополнительными задачами дисциплины и проведения лабораторного практикума являются:

- изучение методов и моделей Data Mining;
- получение представления об алгоритмах построения деревьев

решений;

- изучение алгоритмов классификации и регрессии;
- изучение алгоритмов поиска ассоциативных правил;
- изучение методов кластеризации.

Теоретические знания и практические навыки, полученные обучаемыми при изучении дисциплины, должны быть использованы в процессе изучения последующих дисциплин по учебному плану.

В результате изучения дисциплины обучающийся должен:

а) знать:

- принципы обработки больших массивов данных, способы их представления и хранения;
- основные задачи и методы интеллектуального анализа данных;
- возможности современных и перспективных средств разработки программных продуктов, технических средств.

б) уметь:

- формулировать задачи анализа данных;
- выбирать адекватные алгоритмы их решения;
- выполнять процедуры проектирования хранилищ данных и заполнения готовых хранилищ данными;
- оценивать качество получаемых решений;
- выбирать средства реализации требований к программному обеспечению.

в) владеть:

- технологиями разработки алгоритмов и программными системами анализа данных;
- средствами автоматизации интеллектуального анализа и обработки данных;
- формирование и предоставление отчетности в соответствии с установленными регламентами.

Лабораторная работа 1 Надстройки интеллектуального анализа данных для MicrosoftOffice

Цель: в ходе данной лабораторной работы будет рассмотрен процесс установки пакета надстроек интеллектуального анализа данных для MicrosoftOffice 2007 и начального конфигурирования MicrosoftSQLServer 2008 (2008 R2).

Один из возможных вариантов проведения интеллектуального анализа данных средствами Microsoft SQL Server 2008 - использование надстроек для пакета Microsoft Office 2007.В этом случае источником данных может служить, например, электронная таблица Excel. Данные передаются на SQL Server 2008, там обрабатываются, а результаты возвращаются Excel для отображения.

Для использования подобной "связки", вам должен быть доступен MS SQL Server 2008 в одной из версий, поддерживающих инструменты DataMining (Enterprise, Developer или с некоторыми ограничениями - Standard), MS Office 2007 в версии Professional или более старшей. На момент написания этого материала, надстроек для MS Office 2010 еще не было.Но как отмечается в msdn (http://msdn.microsoft.com/ru-ru/library/bb510513.aspx),32-х разрядная версия Excel 2010 может работать с текущей версией надстроек. В дальнейшем скриншоты будут приводиться именно для сочетания MSOffice 2010 и надстроек интеллектуального анализа для Office 2007.

Сами надстройки интеллектуального анализа данных для MSOffice 2007 свободно доступны на сайте Microsoft по адресу (ссылка приводится для локализованной версии, возможно, выпущены более свежие версии): http://www.microsoft.com/downloads/ru-ru/details.aspx?FamilyID=a42c6fa1-2ee8-43b5-a0e2-cd30d0323ca3&displayLang=ru

Особых сложностей процесс установки надстроек не вызывает. Единственное, что хочется отметить, по умолчанию предлагается устанавливать не все компоненты. Но для выполнения дальнейших лабораторных, лучше сделать полную установку (рис. 4.1)

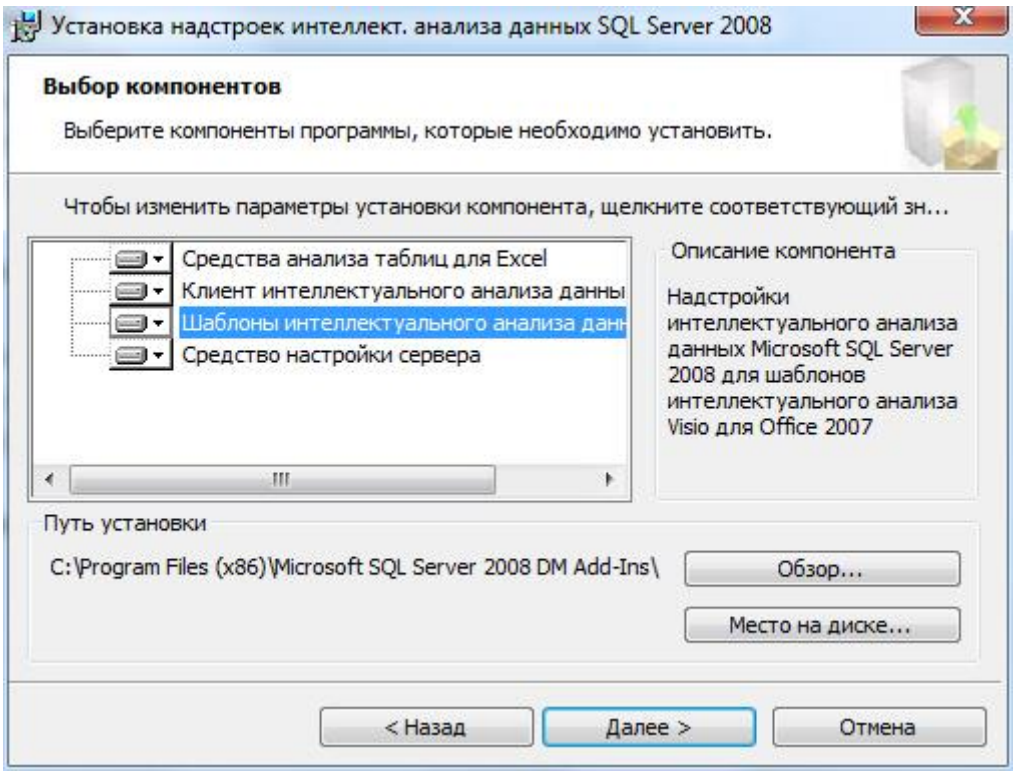


Рис. 4.1. Выбор устанавливаемых компонентов

Следующий шаг - конфигурирование MS SQLServer для работы с надстройками. Для этого используется мастер "Приступая к работе" (GettingStarted), запускаемый из главного меню (рис. 4.2)

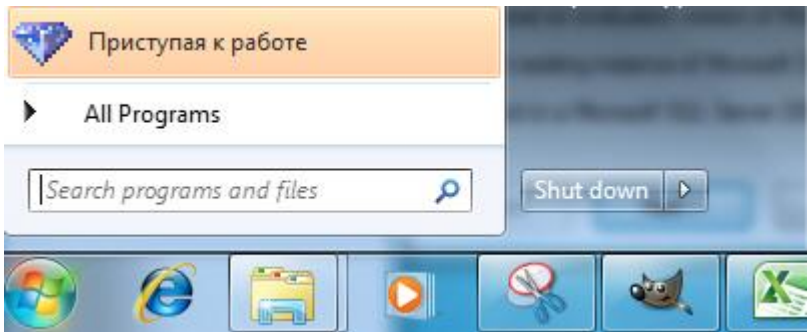


Рис. 4.2. Запуск мастера "Приступая к работе"

Для того, чтобы выполнить конфигурацию MS SQLServer 2008 надо иметь там права администратора. На первом шаге мастер предлагает выбрать, скачать ли пробную версию MS SQLServer, конфигурировать существующий экземпляр сервера, где у пользователя администраторские права, или использовать сервер, на котором пользователь не является администратором (в этом случае, будет сформировано письмо администраторам, с просьбой произвести настройку).

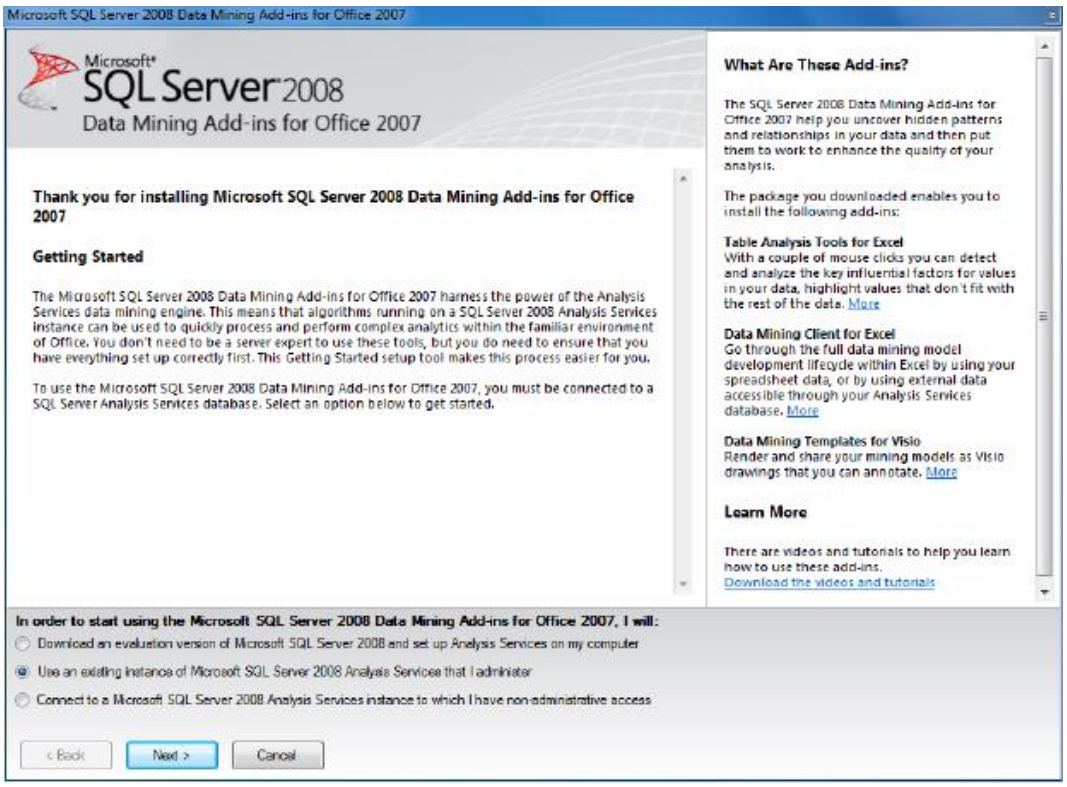


Рис. 4.3. Выбор сервера баз данных

Рассмотрим вариант 2, при выборе которого мастер покажет окно со ссылкой на инструмент "Средство настройки сервера". Его также можно запустить из меню Пуск->Надстройки интеллектуального анализа данных->Средство настройки сервера (рис. 4.4).

Следующее окно предлагает выбрать конфигурируемый сервер(рис. 4.5). По умолчанию стоит "localhost", что соответствует неименованному экземпляру MS SQLServer, установленному на тот же компьютер, на котором запущено "средство настройки". Если это не так, надо указать имя сервера или для именованного экземпляра <имясервера>\<имя экземпляра>.

В окне, представленном на рис. 4.6, дается разрешение на создание временных моделей интеллектуального анализа (Allow creating temporary mining models). Временная модель отличается от постоянной тем, что создается только на время сеанса пользователя. Когда пользователь, проводящий анализ с помощью надстроек, завершит сессию (закроет Excel), модель будет удалена, но результаты анализа сохранятся в электронной таблице. Постоянная модель автоматически не удаляется, хранится на сервере, и к работе с ней можно вернуться.

Рис. 4.4. Средство настройки сервера

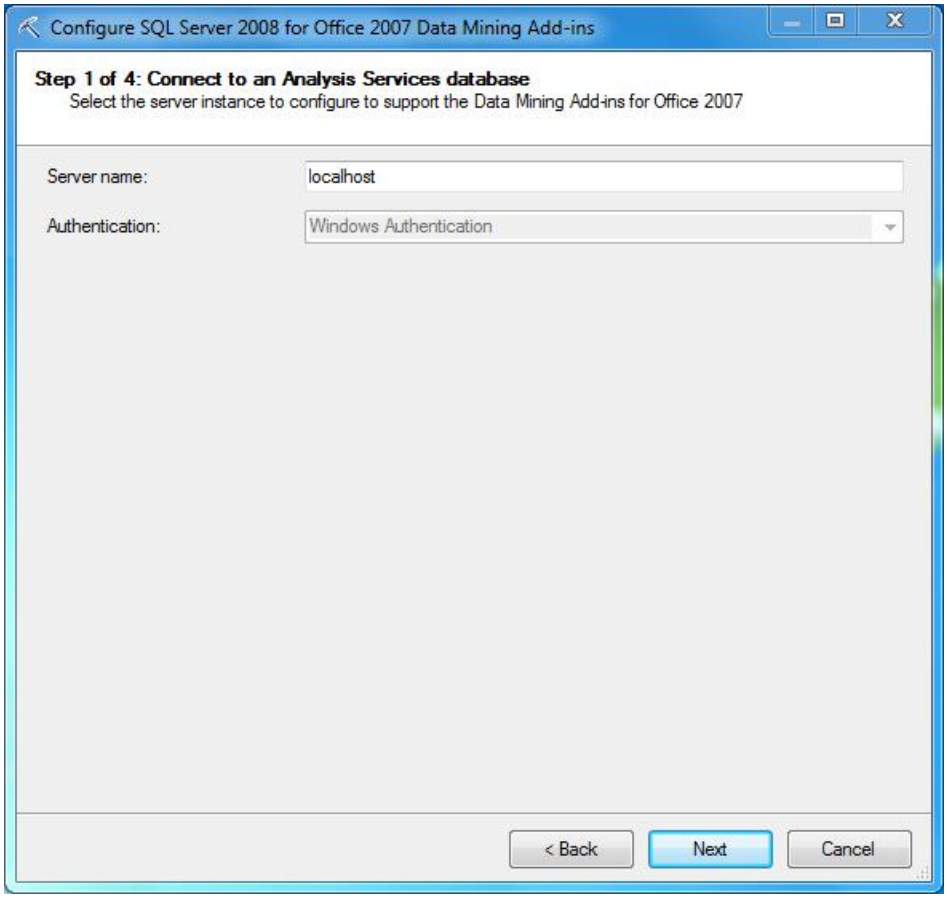


Рис. 4.5. Выбор используемого экземпляра MSSQLServer

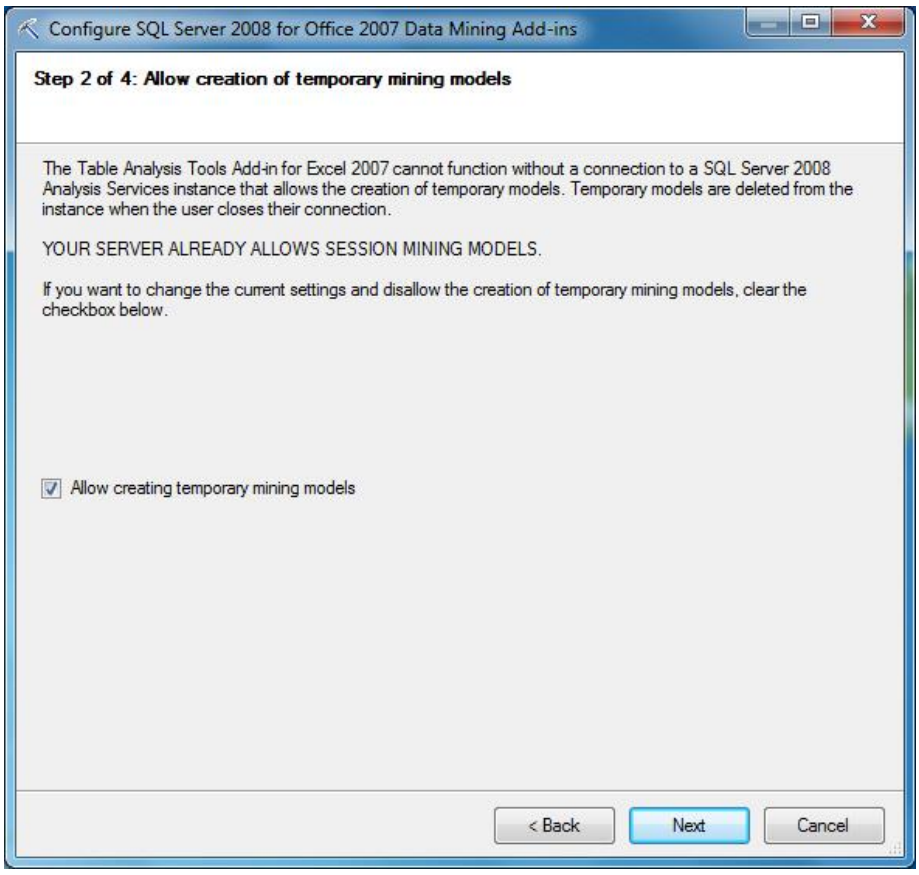


Рис. 4.6. Установка разрешения для создания временных моделей интеллектуального анализа

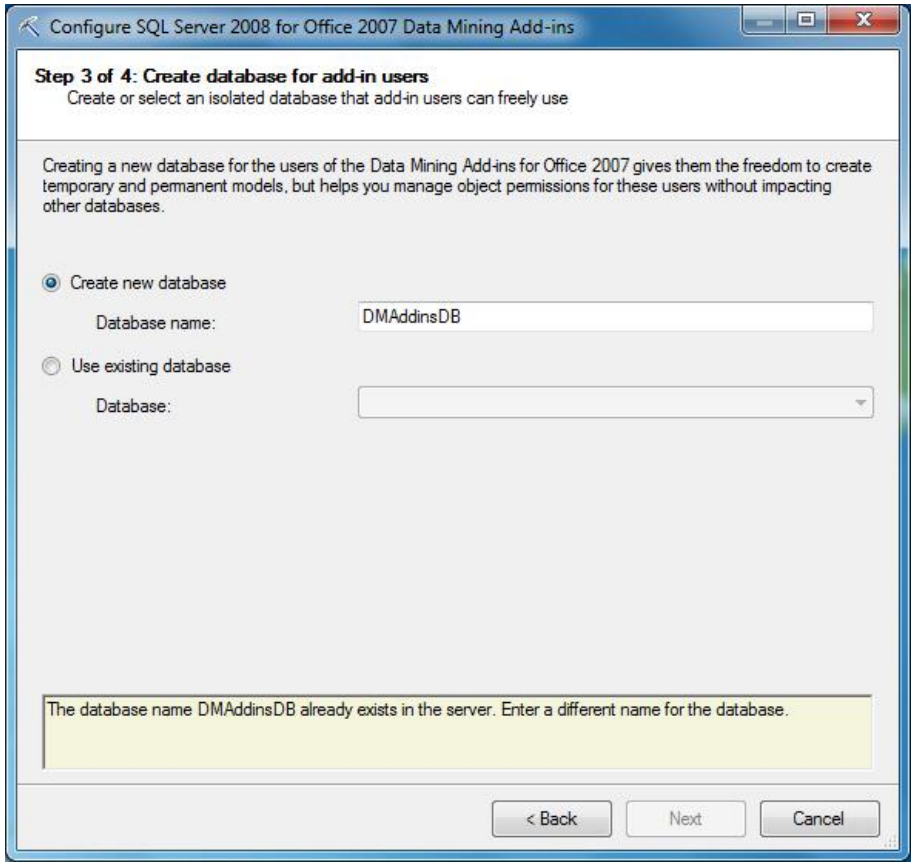


Рис. 4.7. Создание или выбор базы данных аналитических служб

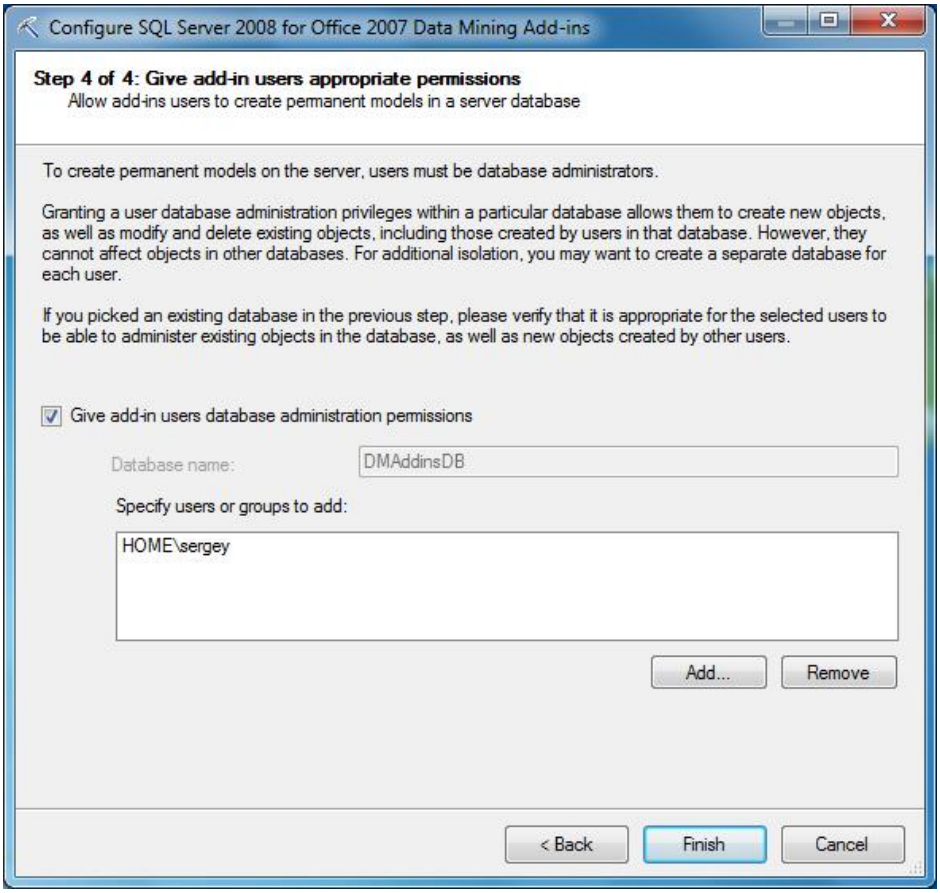


Рис. 4.8. Добавление пользователей в список администраторов выбранной базы

После этого предлагается создать новую базу данных аналитических служб (рис. 4.7) или выбрать для работы существующую.

В окне, представленном на рис. 4.8, можно добавить пользователей в список администраторов созданной базы данных. Это нужно для создания на сервере постоянных моделей. Если использовать только временные модели, права администратора пользователю необязательны.

По окончании настройки можно открыть Excel (а при использовании мастера "Приступая к работе", он будет запущен автоматически с документом "Образцы данных…") и протестировать подключение к серверу. Для этого надо перейти на вкладку DataMining и в разделе Connection (подчеркнут на рис. 4.9) нажать кнопку DMAddinsDB. Появится окно, отображающее настроенные соединения. Кнопка TestConnection позволяет проверить подключение.

Если настроенного соединения нет и кнопка DMAddinsDB выглядит как на рис. 4.11, то нужно создать новое соединение, выбрав в окне Analysis Services Connection(рис. 4.10) кнопку New.

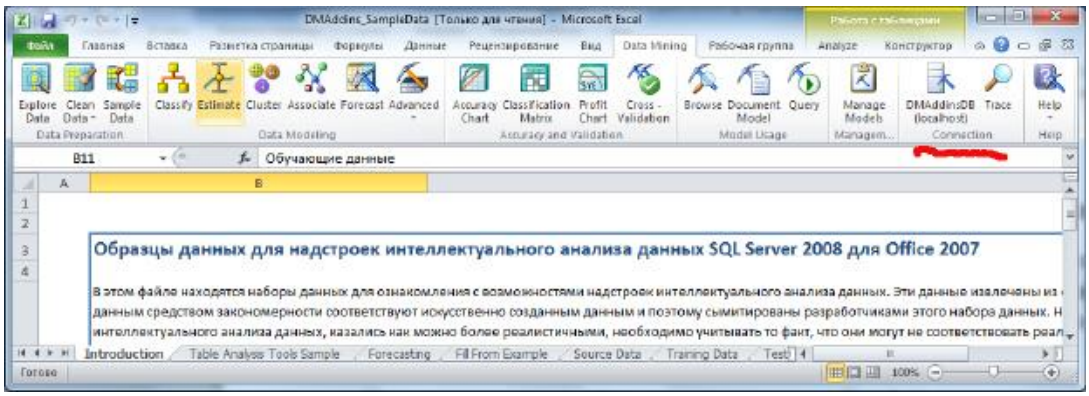


Рис. 4.9. Вкладка DataMining

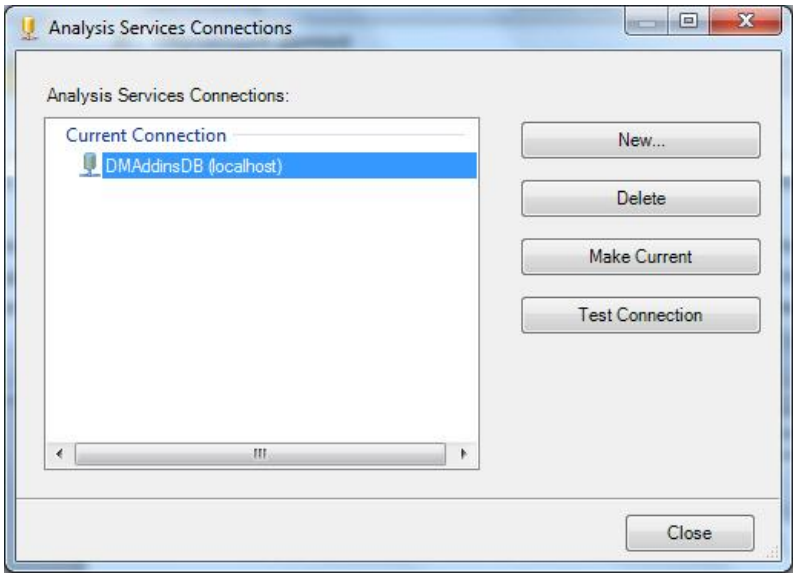


Рис. 4.10. Настроенные соединения

Рис. 4.11. Настроенных соединений нет

При создании нового подключения (рис. 4.12) надо указать сервер, к которому планируете подключаться, и в разделе Catalogname рекомендуется явным образом указать базу данных, с которой будет работать надстройки.

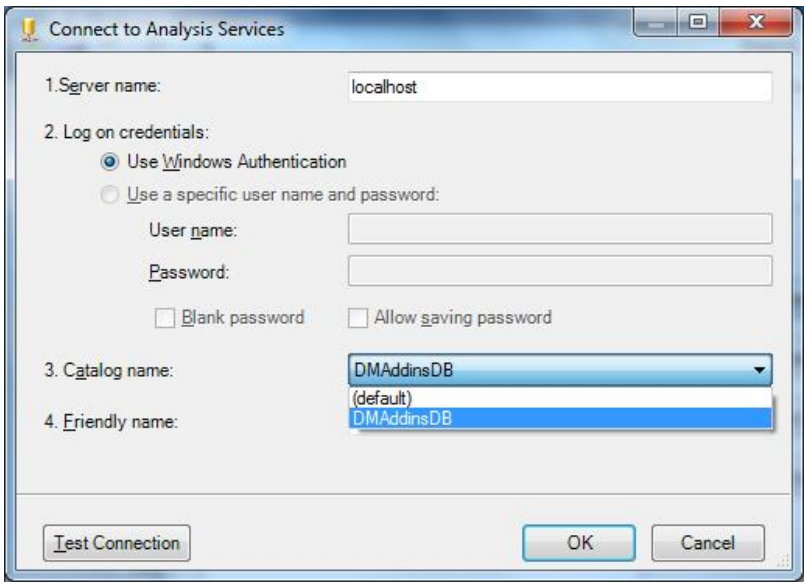


Рис. 4.12. Создание нового подключения

Когда соединение создано и проверено, можно начинать работу. В следующих нескольких лабораторных работах нужно будет использоваться готовый набор данных для анализа. Если же вы планируете работать с собственными данными, необходимо учитывать, что инструменты интеллектуального анализа таблиц работают с данными, отформатированными в виде таблицы. Поэтому ваши данные в Excel нужно выделить и выбрать "Форматировать как таблицу" (рис. 4.13). После этого надо выбрать стиль таблицы и указать заголовок. Вкладка Analyze с инструментами TableAnalysisTools появится при щелчке в области таблицы (рис. 4.14).

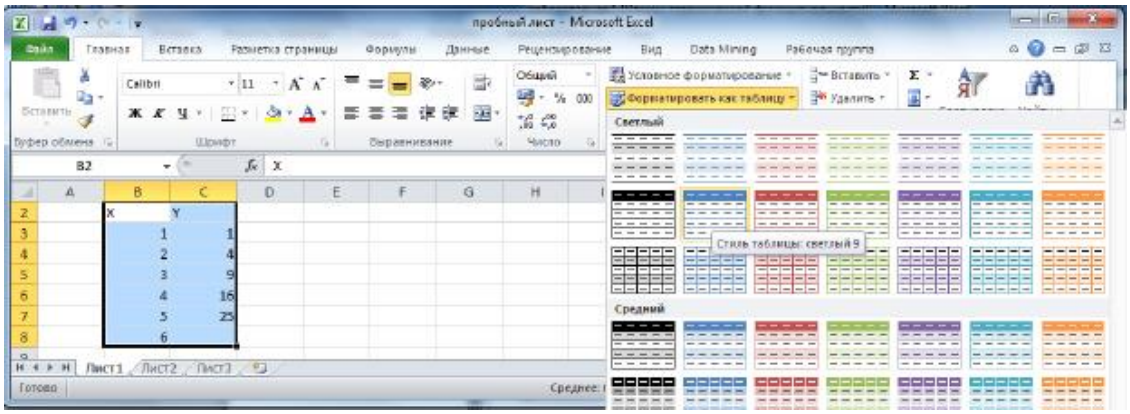


Рис. 4.13. Форматирование подготовленных данных

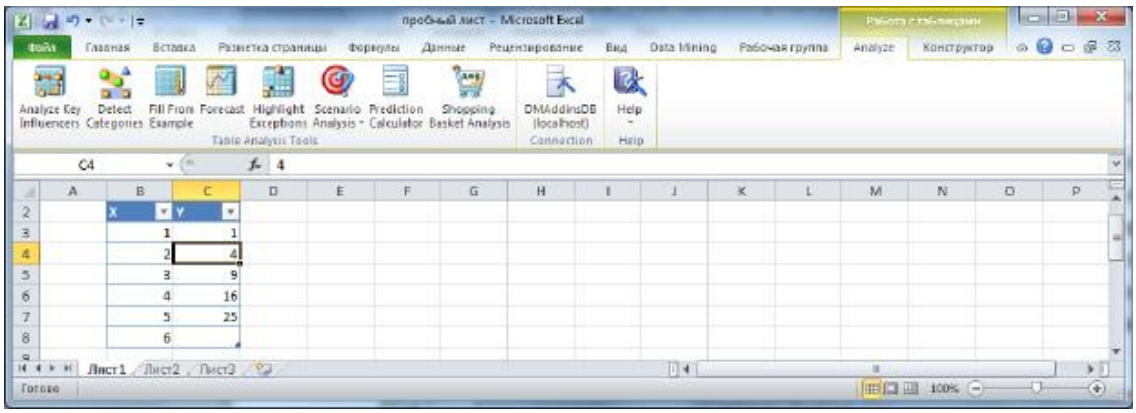


Рис. 4.14. Вкладка с инструментами интеллектуального анализа таблиц

Задание 1. Установите надстройки интеллектуального анализа данных для MicrosoftOffice 2007. Выполните необходимую конфигурацию MSSQLServer 2008 (2008 R2) для работы с надстройками. Создайте и протестируйте подключение.

Задание 2. Подготовленный набор данных (для примера, можно взять приведенный на рис. 4.14) отформатируйте как таблицу. Убедитесь, что вы можете получить доступ к вкладке с инструментами интеллектуального анализа таблиц.

Лабораторная работа 2 Использование инструментов "AnalyzeKeyInfluencers" и "DetectCategories"

Цель: В ходе данной лабораторной работы будет рассмотрено использование инструментов "Анализ ключевых факторов влияния" ("AnalyzeKeyInfluencers") и "Обнаружение категорий" ("DetectCategories"), относящихся к компоненту "Средства анализа таблиц для Excel" пакета надстроек интеллектуального анализа данных для MicrosoftOffice 2007.

Начнем непосредственное изучение инструментов интеллектуального анализа данных (DataMining, сокр.DM). В состав пакета надстроек для MS Office 2007 входит электронная таблица с образцами данных. Она может быть открыта из меню Пуск->Надстройки интеллектуального анализа данных. Microsoft SQL Server 2008. Но переведено содержимое файла только частично - первая страница с оглавлением и некоторые заголовки. Поэтому в работе будет использоваться локализованный набор данных для анализа, доступный для скачивания по адресу http://russiandmaddins.codeplex.com/.

Скачайте файл, откройте его и отформатируйте данные на листе "клиенты" как таблицу (см. "Надстройки интеллектуального анализа данных для MicrosoftOffice"). Перейдите на вкладку Analyze (рис. 5.1). Анализируемая таблица содержит данные фирмы, продающей велосипеды. В ней собрана информация о клиентах (идентификатор, семейное положение, пол и т.д.) и указано, приобрел клиент велосипед или нет.

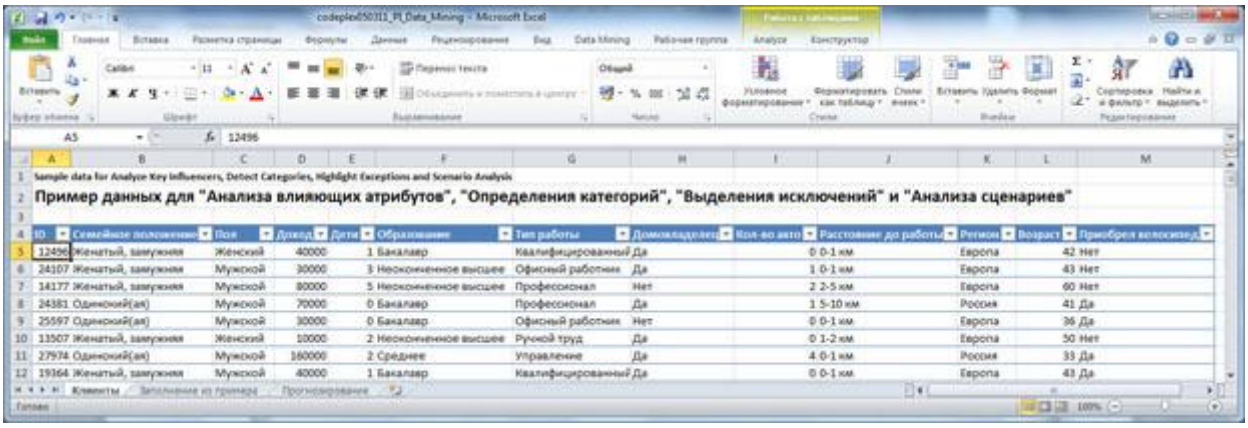


Рис. 5.1. Подготовленный набор данных

Анализ ключевых факторов влияния

Инструмент AnalyzeKeyInfluencers позволяет определить, как зависит интересующий нас параметр от других. При этом важно правильно определить, что и от чего может зависеть. Собственно, в этом отчасти и заключается мастерство аналитика, основанное на его знании предметной области и используемых методов DM.

В связи с тем, что мы оцениваем степень взаимного влияния разных параметров друг на друга, стоит сразу убрать из рассмотрения полностью независимые и наоборот, полностью зависимые. Пусть, например, мы хотим

оценить влияние различных факторов на уровень заработной платы человека. Если у нас есть поле, содержащее уникальный идентификатор (например, порядковый номер записи в таблицы или номер паспорта), его стоит убрать из рассмотрения, как не влияющий на значение исследуемого параметра. Другой пример, пусть у нас есть значениезаработной платы за месяц и за год, рассчитываемое как заработная плата за месяц, умноженная на 12. Мы знаем, что эти значения всегда связаны, искать зависимость одного от другого средства DM не имеет смысла, а имеющаяся сильная зависимость скроет влияние других факторов, которое мы как раз и хотим выявить.

Теперь определим, от чего зависит решение клиента о покупке велосипеда. Нажимаем на кнопку Analyze Key Influencers и указываем в качестве целевого столбца столбец "Приобрел велосипед" (рис. 5.2). Перейдем по ссылке "Choose columns to be used for analysis", чтобы указать параметры, влияние которых мы хотим оценить (рис. 5.3). Здесь сбросим отметку напротив "ID" и "Приобрел велосипед"(хотя последнее можно и не делать).

Рис. 5.2. Выбор зависимого параметра для анализа

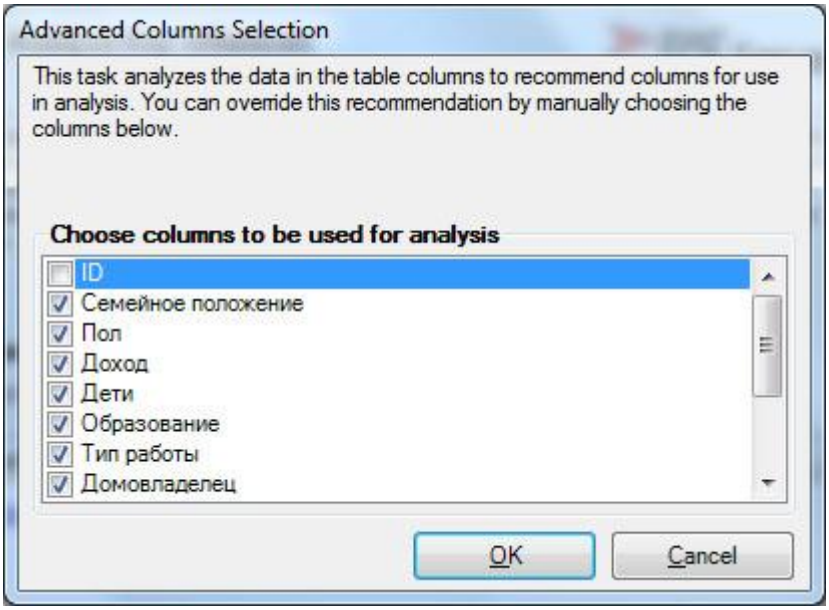


Рис. 5.3. Выбор параметров, от которых зависит анализируемый

После запуска процедуры анализа (по кнопке Run, рис. 5.2) будет сформирован отчет о факторах влияния и предложено формирования дополнительного сравнительного отчета (рис. 5.4). В основном отчете указывается столбец (Column), его значение (Value), значениецелевого столбца, с которым оно связывается (Favors) и уровень влияния (Relative Impact), оцениваемый по шкале от 0 до 100 балов. Из представленногона рис. 5.4 отчета видно, что на решение не покупать велосипед в наибольшей степени влияет наличие 2-х автомобилей. В то же время не следует воспринимать оценку 100 баллов, как признак того, что в 100% случаев владельцы 2-х машин велосипед не покупали (посмотрите набор данных, там есть и сочетания "2 машины - велосипед куплен", но их меньшинство). Второй по уровню влияния на отказ от покупки фактор - "Семейное положение"="женатый, замужем".

Наибольшее влияние на положительное решение о приобретении велосипеда оказывает отсутствие у клиента машины.

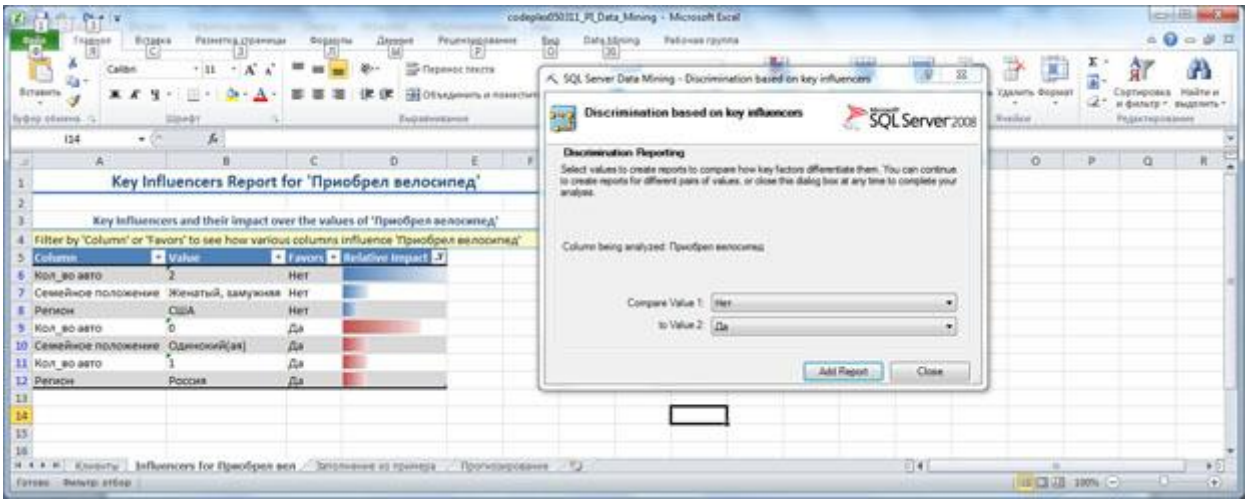


Рис. 5.4. Основной отчет

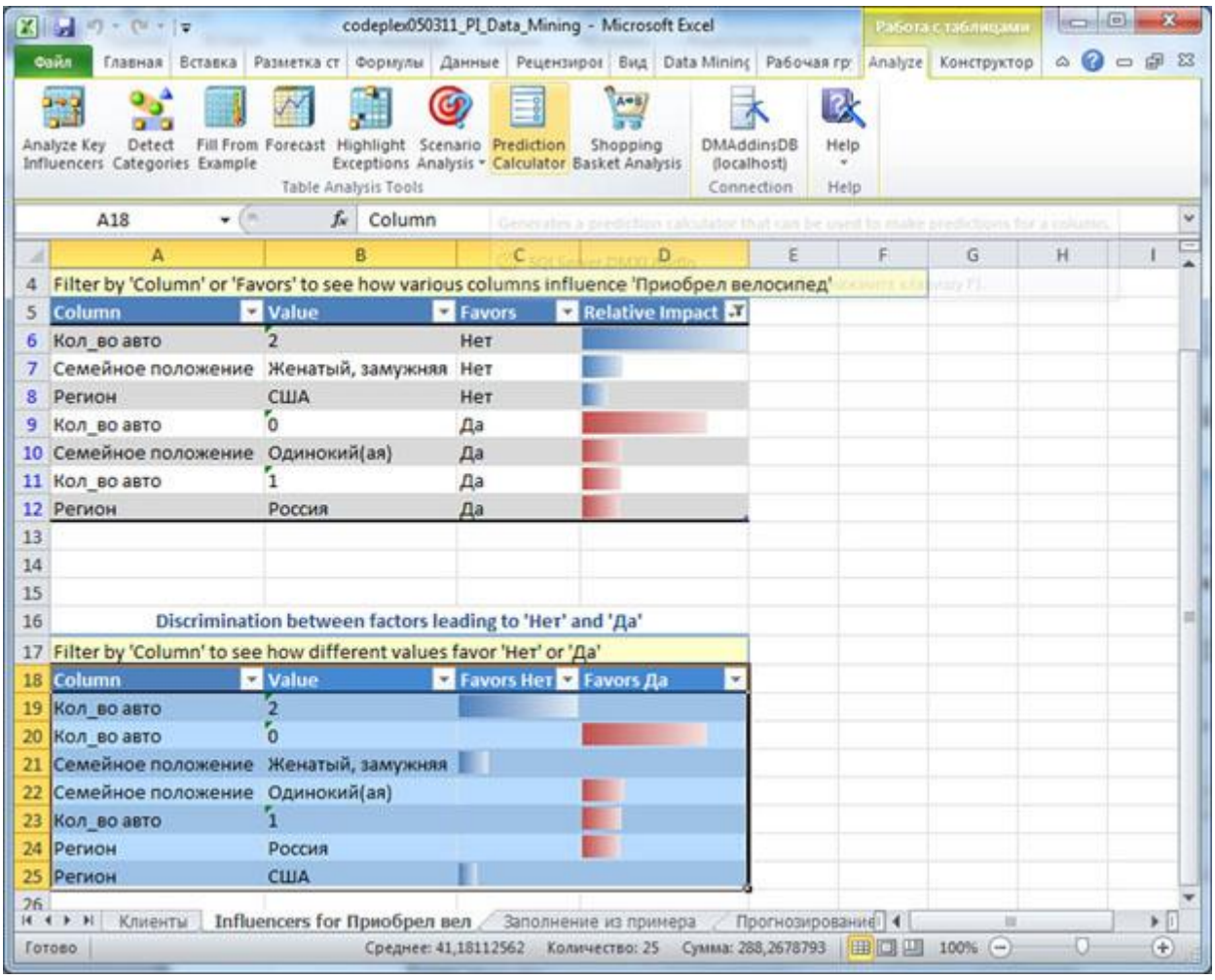


Рис. 5.5. Сравнительный отчет

Если добавить сравнительный отчет для двух выбранных значений (рис. 5.4, Add Report), можно увидеть, чем отличается выбор в пользу одного значения целевого столбца от выбора в пользу другого (рис. 5.5). В нашем примере просто произойдет перегруппировка исходного отчета, т.к. возможных значений всего 2. В других случаях, дополнительный отчет позволяет провести детальное сравнение двух выбранных вариантов.

Как отмечается в [1], если целевой или другой столбец, обрабатываемый инструментом Analyze Key Influencers,содержит много различных числовых значений, то проводится дискретизация. Весь интервал значений делится на несколько диапазонов, каждый из которых рассматривается как одно из возможных значений (например, вместо точного значения 2,5 мы получим "диапазон от 2 до 3").

Задание 1. Проведите анализ в соответствии с рассмотренным примером.

Задание 2. На том же наборе данных проанализируйте зависимость уровня дохода от образования, семейного положения, типа работы, пола, возраста и региона проживания клиента. Опишите результаты.

Дополните отчет сравнительным анализом для самого низкого и следующего за ним диапазона дохода. А затем - для самого низкого и самого

высокого диапазона. Опишите результаты проведенного анализа и предложите их интерпретацию.

Задание 3. Предложите свой вариант анализа данных, и пример использования полученных результатов.

Сформированный отчет будет доступен и в случае, если вы откроете файл и на другом компьютере (без подключения к аналитическим службам SQLServer).

Чтобы вернуть данные в исходное состояние нужно удалить листы с сформированными отчетами.

Обнаружение категорий

Инструмент Detect Categories позволяет решить задачу кластеризации, т.е. разделения всего множества вариантов на "естественные" группы, члены которых наиболее близки по ряду признаков. Подобная задача также называется задачей сегментации.

Итак, в нашем примере есть описание множества клиентов и нужно разделить их на небольшое количество групп (чтобы отдельным группам сформировать специальное предложение и т.п.).

В связи с тем, что в процессе работы инструмент добавляет данные в исходную таблицу, рекомендуется перед началом работы сделать ее копию (рис. 5.6).

После этого нажимаем кнопку Detect Categories и настраиваем параметры (рис. 5.7). Здесь хочется обратить внимание на атрибут ID, который как было отмечено выше, не имеет смысл учитывать в ходе анализа. Поэтому он автоматически исключен. В нашем случае, остальные атрибуты можно оставить. Еще раз хотелось бы повторить, что этот выбор каждый раз делается исходя из особенностей предметной области.

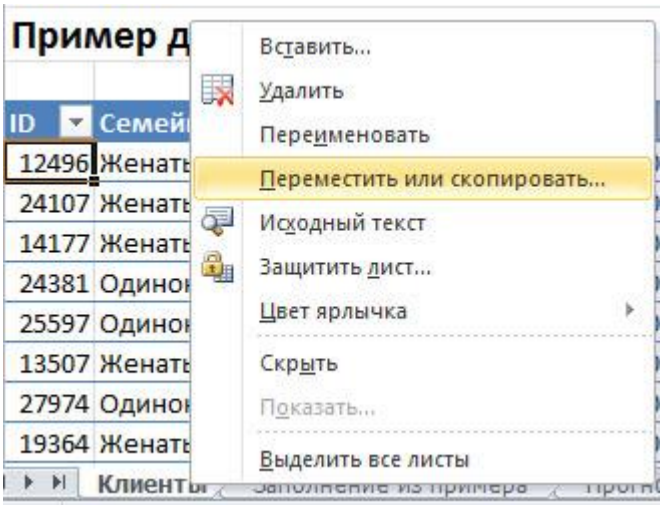


Рис. 5.6. Перед началом работы лучше скопировать лист Excel

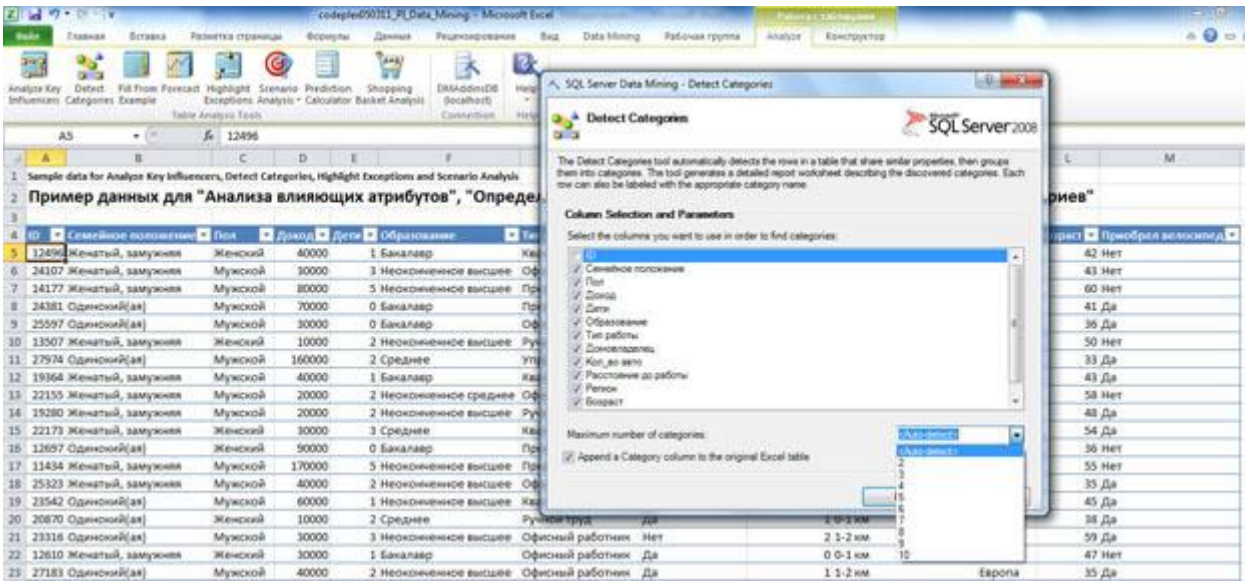
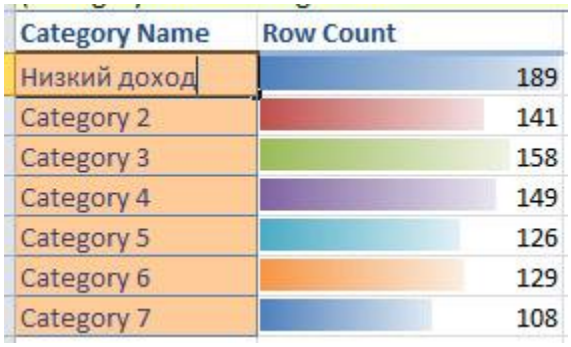


Рис. 5.7. Выбор параметров, которые будут анализироваться

Кроме указания учитываемых параметров, можно явно указать число категорий (или оставить по умолчанию автоматическое определение). Также по умолчанию поставлен флажок "Appenda Category column to the original Excel table", указывающий, что к записям в исходной таблице будет добавлено указание на категорию.

Сформированный отчет содержит 3 раздела. В первом указаны определенные инструментом категории и число строк, попадающих в каждую из них (рис. 5.8). Поле с названием категории допускает редактирование и можно сопоставить категории более значимое название. Например, как будет показано ниже, для клиентов первой категории характерен низкий доход и ее можно так и назвать. Когда мы введем это название, везде кроме диаграммы Category Profiles Chat, оно автоматически заменит "Category 1" (чтобы название поменять и на диаграмме, надо нажать <Alt>+<Ctrl>+<F5>).

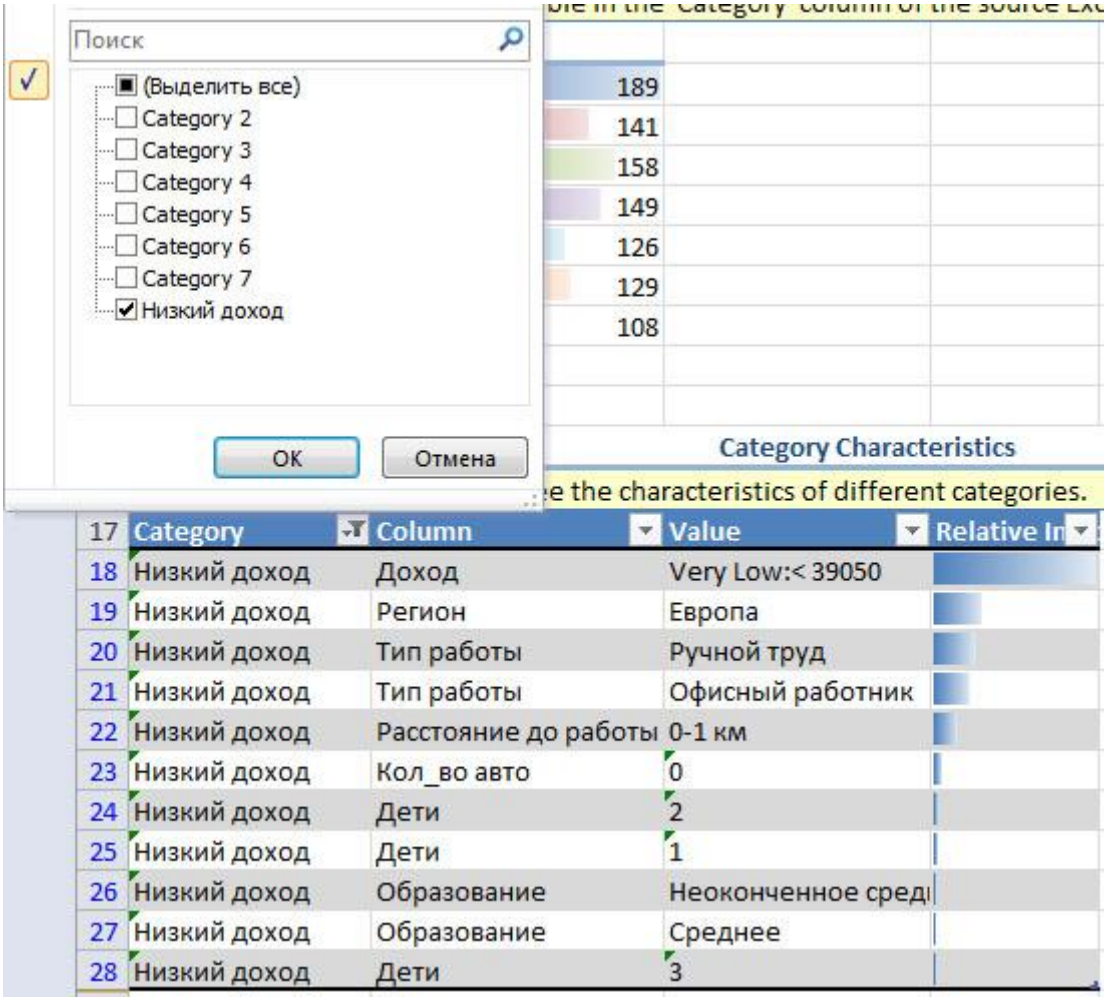


Category Name	Row Count
Низкий доход	189
Category 2	141
Category 3	158
Category 4	149
Category 5	126
Category 6	129
Category 7	108

Рис. 5.8. Выделенные категории

Следующий раздел отчета описывает характеристики выделенных категорий и степень влияния каждого параметра (рис. 5.9). По умолчанию отображается информация только по одной категории, но щелчком мыши по иконке фильтра на заголовке таблицы можно установить отображение всех

категорий или какого-то их сочетания, как это показано на рисунке.



17	Category	Column	Value	Relative In
18	Низкий доход	Доход	Very Low:< 39050	
19	Низкий доход	Регион	Европа	
20	Низкий доход	Тип работы	Ручной труд	
21	Низкий доход	Тип работы	Офисный работник	
22	Низкий доход	Расстояние до работы	0-1 км	
23	Низкий доход	Кол_во авто	0	
24	Низкий доход	Дети	2	
25	Низкий доход	Дети	1	
26	Низкий доход	Образование	Неоконченное сред	
27	Низкий доход	Образование	Среднее	
28	Низкий доход	Дети	3	

Рис. 5.9. Описание категории

Третий раздел отчета - это диаграмма профилей категорий. Она показывает количество строк данных в каждой категории с каждым значением выбранных параметров. По умолчанию отображается только один параметр. Для рассматриваемого примера это возраст. Но в нижней части диаграммы есть фильтр Column, с помощью которого можно изменить число параметров. Например, на рис. 5.10 для каждой категории отображается распределение по возрасту и доходу. Из него видно, что клиенты переименованной нами категории "Низкий доход" на самом деле имеют очень низкий доход. А клиенты категории 3 в подавляющем большинстве очень молоды.

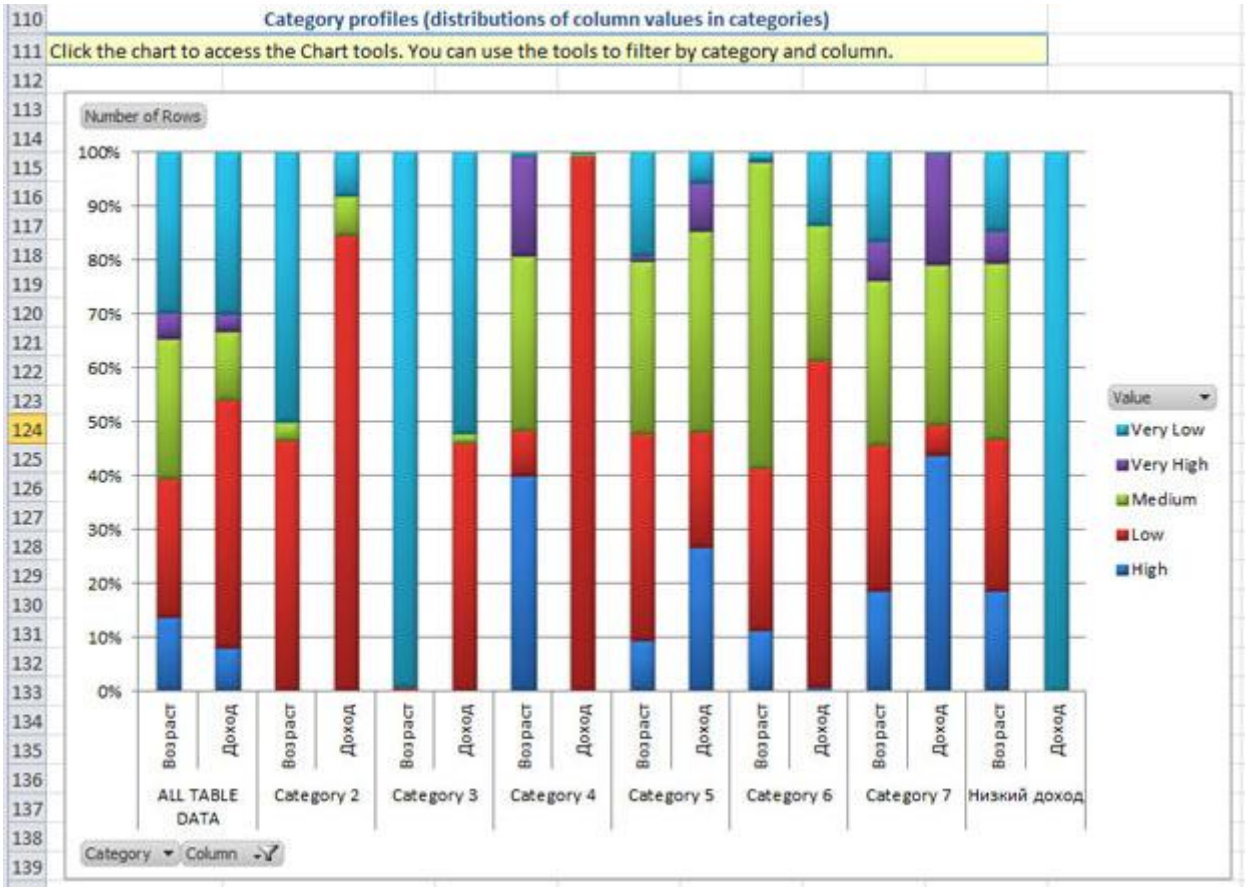


Рис. 5.10. Диаграмма профилей категорий

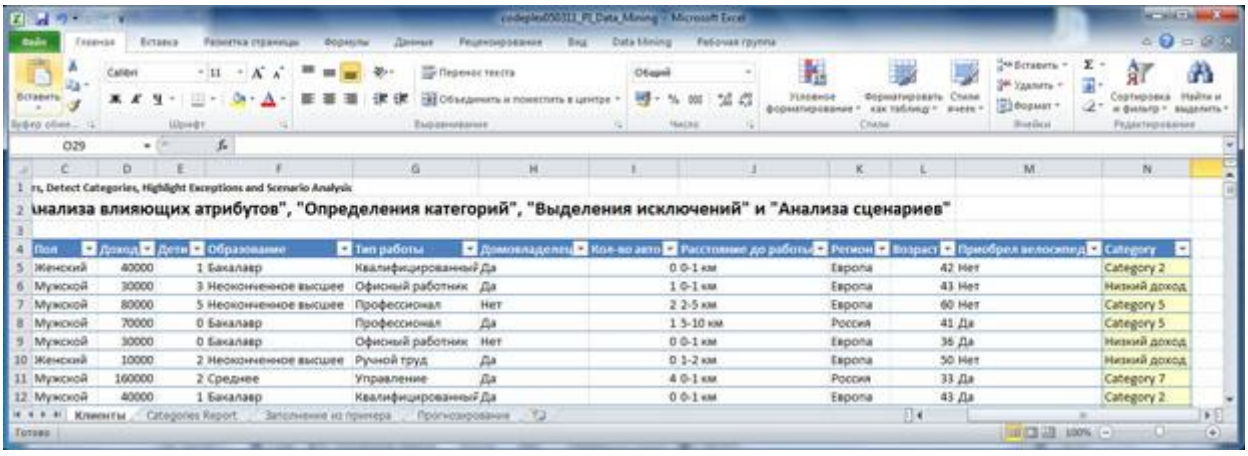


Рис. 5.11. Сопоставление категорий записям в исходной таблице

Рисунок 5.11 показывает, что всем записям исходной таблицы теперь сопоставлена категория, к которой они относятся. А с помощью фильтров можно просмотреть записи, относящиеся к выбранной категории.

Задание 1. Переименуйте категорию Category 3.

Задание 2. Проведите анализ параметров, характеризующих оставшиеся категории, и дайте им осмысленные названия.

Лабораторная работа 3. Использование инструментов "FillFromExample" и "Forecast"

Цель: В данной лабораторной работе будет рассмотрено использование инструментов "Заполнение по примеру" ("FillFromExample") и "Прогноз" ("Forecast"), относящихся к компоненту "Средства анализа таблиц для Excel" пакета надстроек интеллектуального анализа данных для MicrosoftOffice 2007.

Оба рассматриваемых инструмента используются для решения задач прогнозирования неизвестных значений параметров. Поэтому в обоих случаях требуется обучающий набор данных, на базе которого строится модель, применяемая для предсказания.

Заполнение по примеру

В качестве учебного набора данных, как и в прошлой лабораторной будем использовать локализованный пример для Excel, взятый с http://russiandmaddins.codeplex.com/

Нужные данные находятся на листе "Заполнение из примера" (рис. 6.1). Здесь описывается ряд клиентов магазина. Для некоторых из них отмечено, является ли данный клиент высокодоходным. Эти строки будут использоваться как обучающая выборка. Задачей анализа будет являться оценка остальных клиентов по этому параметру.

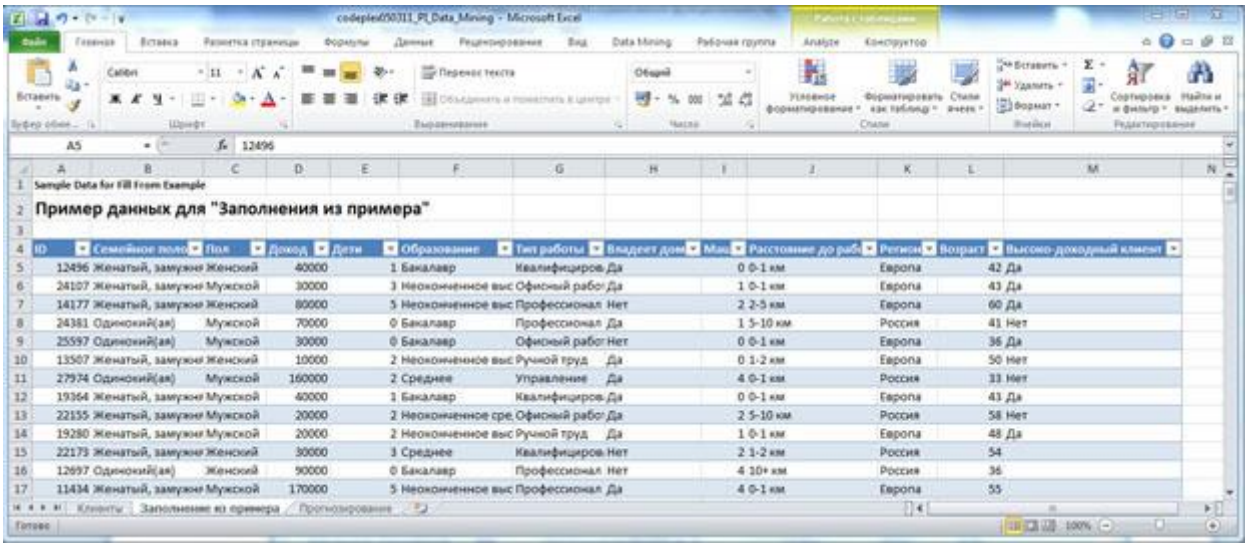


Рис. 6.1. Набор данных для инструмента FillFromExample

Для решения этой задачи используется алгоритм MicrosoftLogisticRegression. Необходимо понимать, что для создания модели в обучающей выборке должны быть представлены варианты со всеми возможными значениями целевого столбца. Необходимое число примеров зависит от особенностей предметной области. Но во многих случаях справедливо, что чем больше характерных примеров в обучающей выборке, тем более качественно будет обучена модель.

Соответственно, данный инструмент непригоден для задачи предсказания значений параметра, который может принимать непрерывные числовые значения.

Еще одна особенность - анализ проводится по столбцам (т.е. предсказывается значение столбца). Если ряд, который необходимо заполнить, хранится в виде строки, перед началом анализа надо выполнить транспонирование (скопировать в буфер, выбрать в контекстном меню "Специальная вставка" и отметить флажок "Транспонировать").

Запустим инструмент FillFromExample. В первом окне будет предложено выбрать столбец, содержащий образцы данных. В нашем случае он автоматически определен верно - "Высокодоходный клиент". Как и в предыдущих случаях, по ссылке "Choosecolumnstobeusedforanalysis", можно выбрать столбцы, учитываемые при анализе. Эвристический механизм определил, что поле ID учитывать не надо. На практике, рекомендуемые настройки стоит менять только в случае, если точно известно о взаимной независимости параметров. После запуска, инструмент формирует отчет об обнаруженных шаблонах (рис. 6.3), и добавляет столбец с предсказанными значениями к исходной таблице.

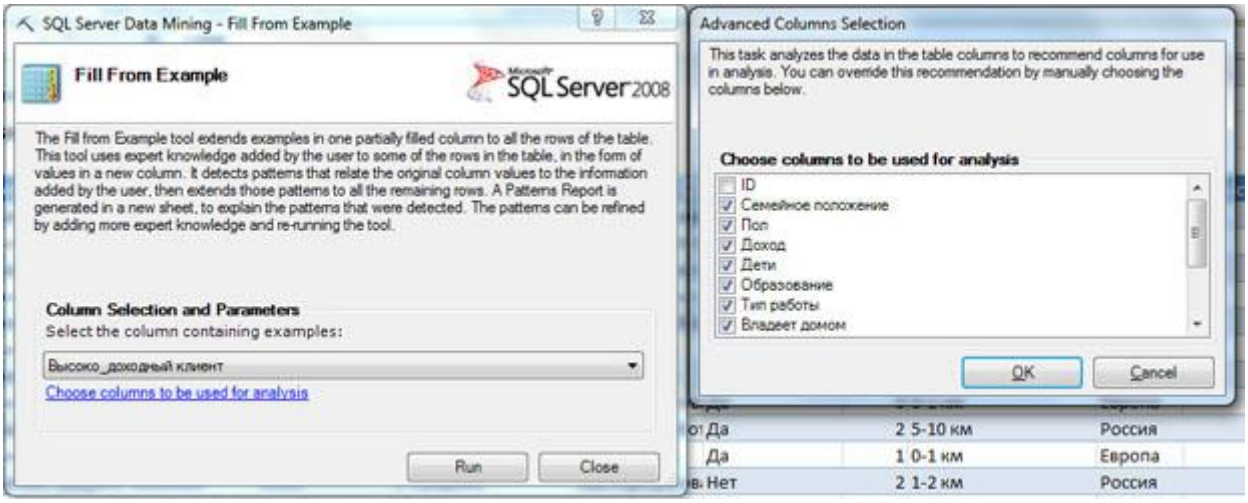


Рис. 6.2. Настройка инструмента FillFromExample

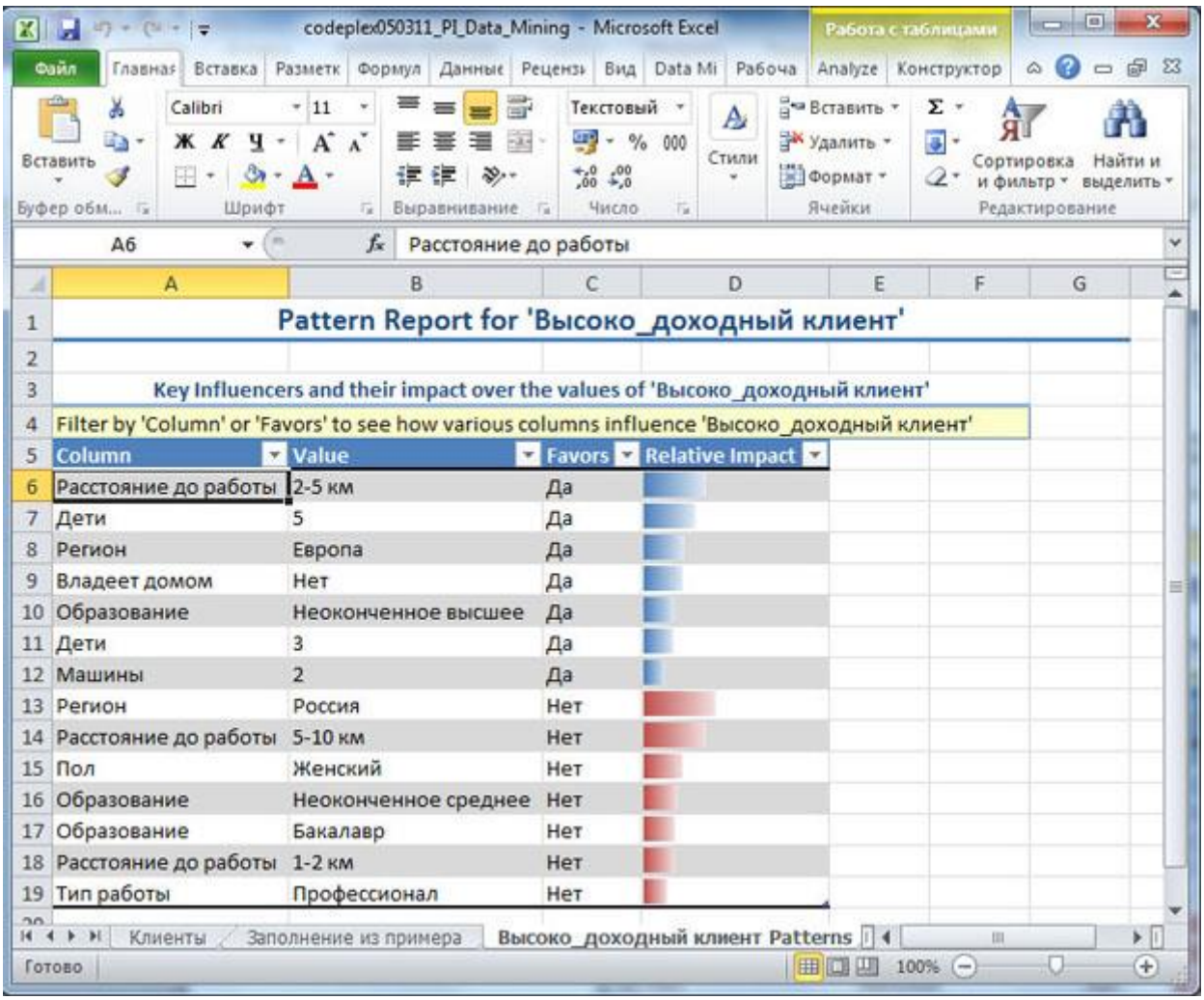


Рис. 6.3. Отчет об обнаруженных шаблонах

В отчете описываются выявленные зависимости между значением целевого столбца (в нашем случае "да" или "нет") и значениями других столбцов. На рис. 6.3 видно, что весовой коэффициент для "Да", соответствующий значению "2-5 км" параметра "Расстояние до работы", равен 34. Это значение имеет самый большой удельный вес при выборе варианта "Да". Это можно интерпретировать, как "расстояние 2-5 км до работы" во многом определяет выбор в пользу покупки велосипеда.

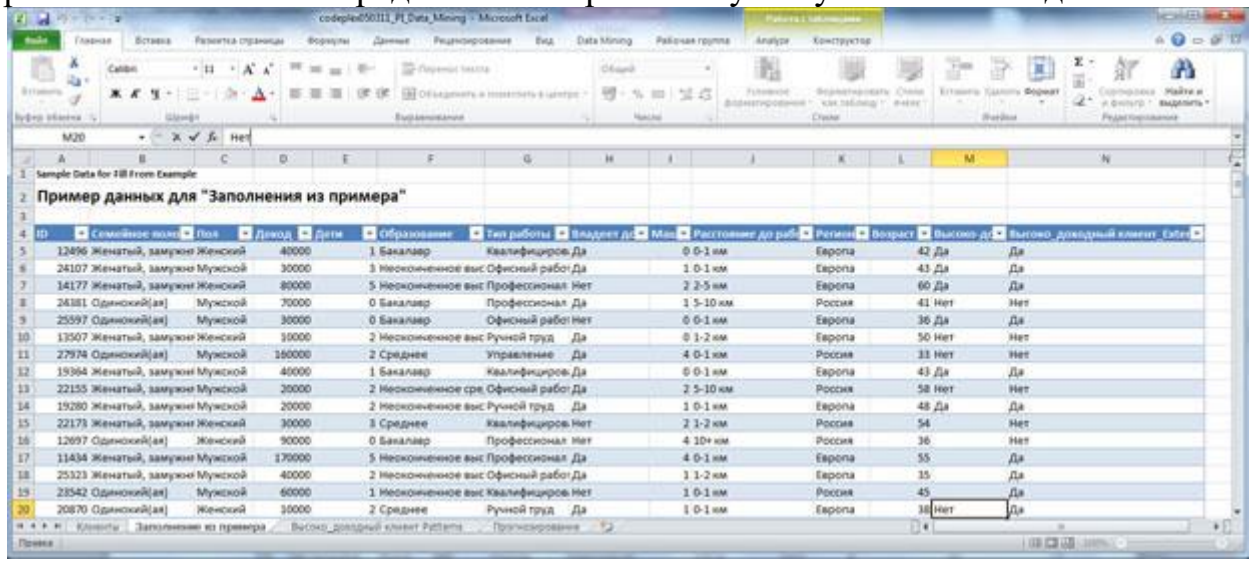


Рис. 6.4. Полученные оценки заносятся в исходную таблицу

Для каждой строки рассчитывается итоговая оценка для каждого варианта (в примере для "Да" и "Нет") и делается выбор в пользу значения с наибольшим суммарным удельным весом. Оно заносится в столбец с суффиксом "_Extended"(на рисунке "Высоко_доходный клиент_Extended"). Для записей, на которых модель обучалась, значение этого столбца совпадает с образцом.

Предположим, мы получили дополнительные данные о каких-то клиентах. Можно изменить образец (рис. 6.4, последняя строка) и снова запустить инструмент. Новые значения будут получены с учетом уточнений в наборе обучающих данных. Подобные итерации позволяют последовательно уточнять производимую оценку значений.

Задание. Проведите анализ и опишите полученные результаты.

Измените обучающий набор данных следующим образом. Найдите строку со значением "расстояние до работы 2-5 км", (например, строку с идентификатором 19562, 97-я строка в таблице) и для параметра "Высоко-доходный клиент" поставьте значение "Нет".Повторите анализ. Как изменился отчет о шаблонах? Объясните эти изменения.

Для того, чтобы полностью удалить результаты работы инструмента, достаточно удалить лист с отчетом и добавленный столбец в таблице с исходными данными.

Прогноз

Инструмент Forecast позволяет построить прогноз значений числового ряда. Ряд должен быть представлен столбцом в таблице (если исследуемые значения организованы в виде строки, требуется, как и в случае инструмента "FillFromExample", выполнить транспонирование).

В используемом нами файле Excel на листе прогнозирование есть набор данных по суммам продаж велосипедов марки M200 по месяцам в трех разных регионах. Таким образом, для исследования мы имеем три числовые последовательности, возможно связанные между собой (рис. 6.5). В процессе работы инструмент строит модель с использованием алгоритма временных рядов (MicrosoftTimeSeries). Для его работы необходимо, чтобы в исследуемых столбцах были только числовые значения (пропуски допустимы). Предсказывать можно числовые (непрерывные) или "денежные" (тип currency) значения. Инструмент не рассчитан на предсказание дат.

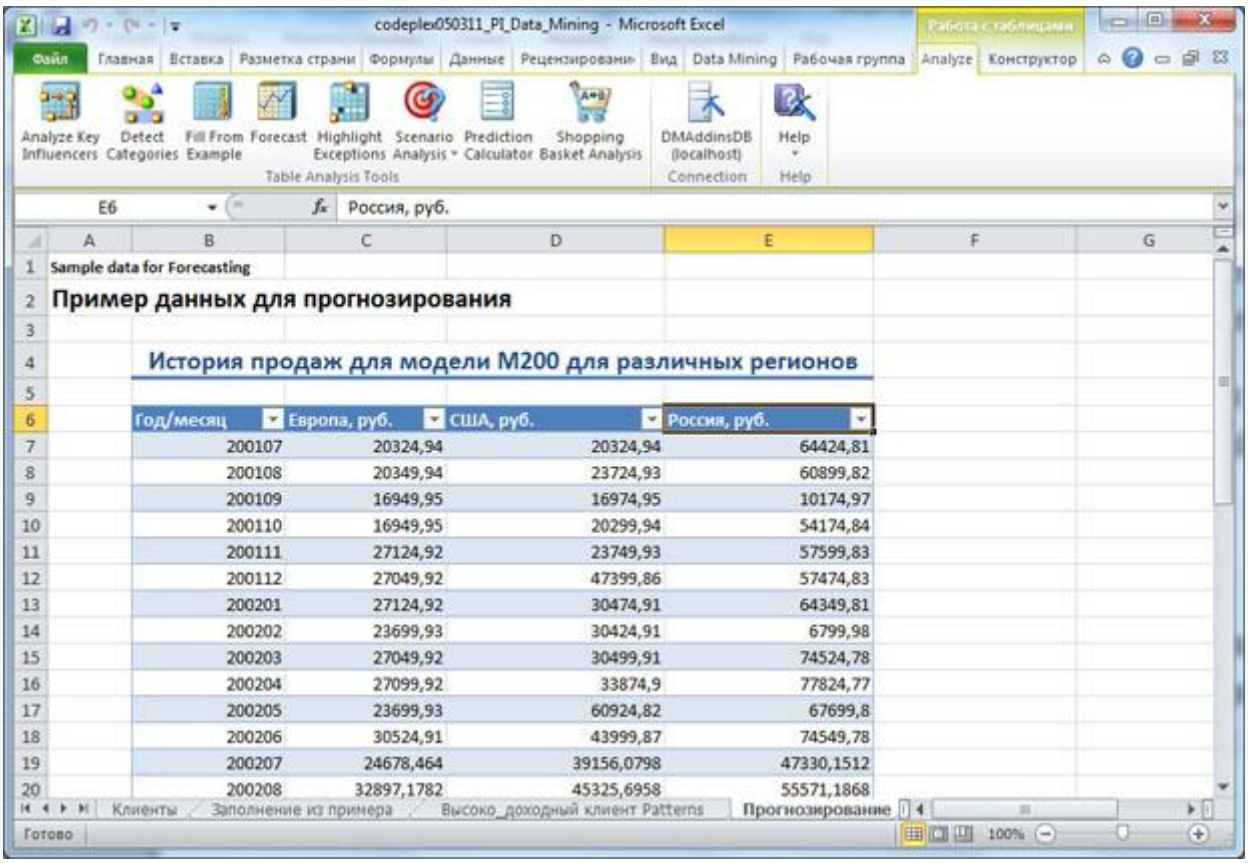


Рис. 6.5. Образец данных для прогнозирования - продажи по месяцам в разных регионах

Как отмечается в [1], инструмент ищет в анализируемой последовательности шаблоны следующих типов:

- тренд - тенденцию изменения значений. Тренд может быть восходящим (возрастание значений ряда) или нисходящим (уменьшение значений);
- периодичность (сезонность) - событие повторяется через определённые интервалы;
- взаимная корреляция - зависимость значений одного ряда от других (например, стоимость акций нефтяных компаний от цен на нефть). Алгоритмы, обнаруживающие взаимную корреляцию, входят в поставку MS SQLServer 2008 версии Enterprise или Developer, а в версии Standard недоступны.

Настройка параметров заключается в выборе анализируемых столбцов, количества предсказываемых значений ряда, указания временной отметки и типа периодичности.

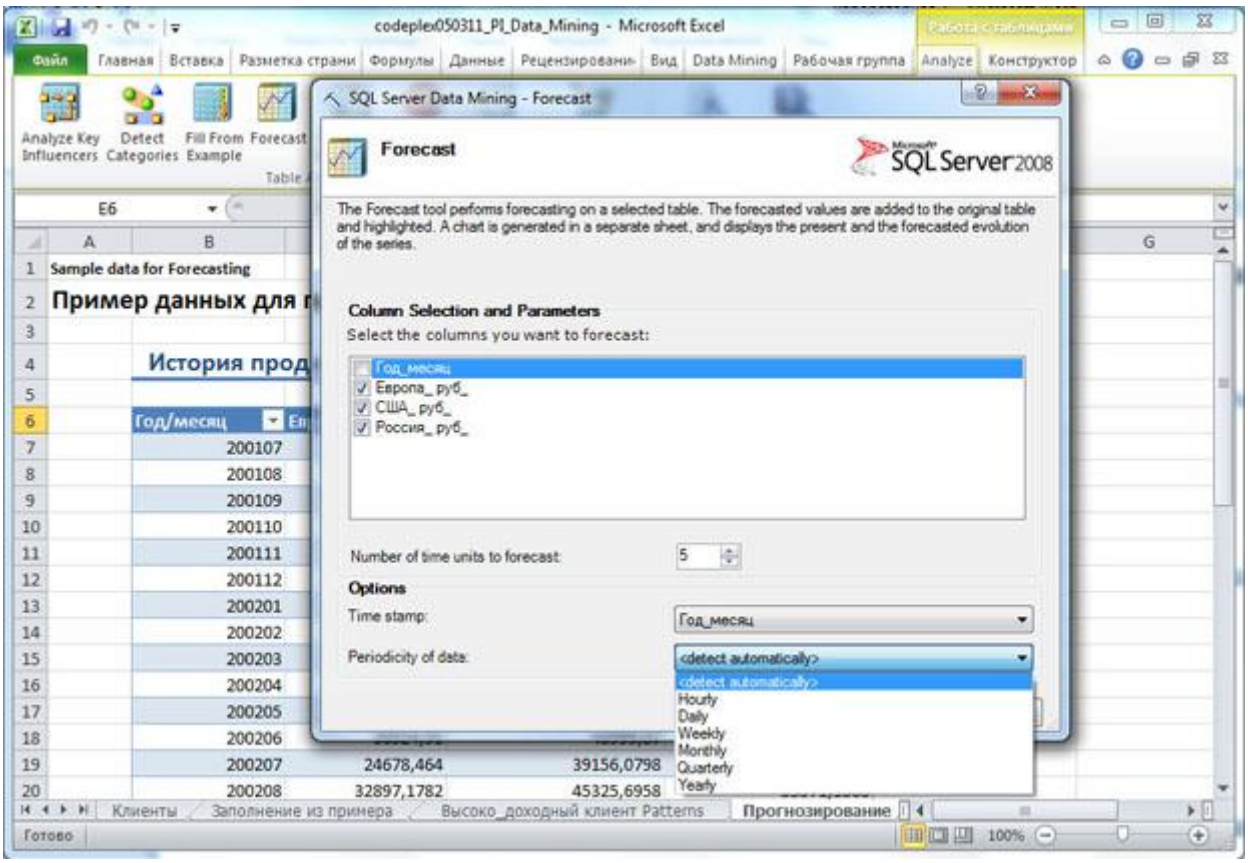


Рис. 6.6. Настройка параметров инструмента Forecast

В нашем случае в качестве временной отметки логично выбрать поле "Год/месяц" (инструмент изменил его название на "Год_месяц" для совместимости с требованиями SQLServer) и согласиться с исключением его из списка предсказываемых. Надо отметить, что значения в столбце, используемом в качестве временной метки, должны быть уникальны.

Что касается периодичности, то предлагаемые для выбора варианты определяются следующим образом[1]:

- Hourly (почасовая) - ищется периодичность 12;
- Daily (дневная) -ищется периодичность 5 и 7 (рабочие дни и неделя полностью);
- Weekly(недельная) - 4 и 13 (число недель в месяце и квартале);
- Monthly (месячная) - 12 (число месяцев в году);
- Yearly - инструмент будет автоматически обнаруживать периодичности.

Если периодичность неизвестна, то рекомендуется оставить "detectautomatically", чтобы инструмент проверил данные на наличие периодичности разных типов.

Инструмент создает отчет с графиком (рис. 6.7), на котором непрерывной линей обозначен "исторический тренд", построенный по имеющимся значениям. Пунктирной линией показано предсказываемое продолжение тренда. Обратите внимание, что временные метки для спрогнозированных значений не проставлены.

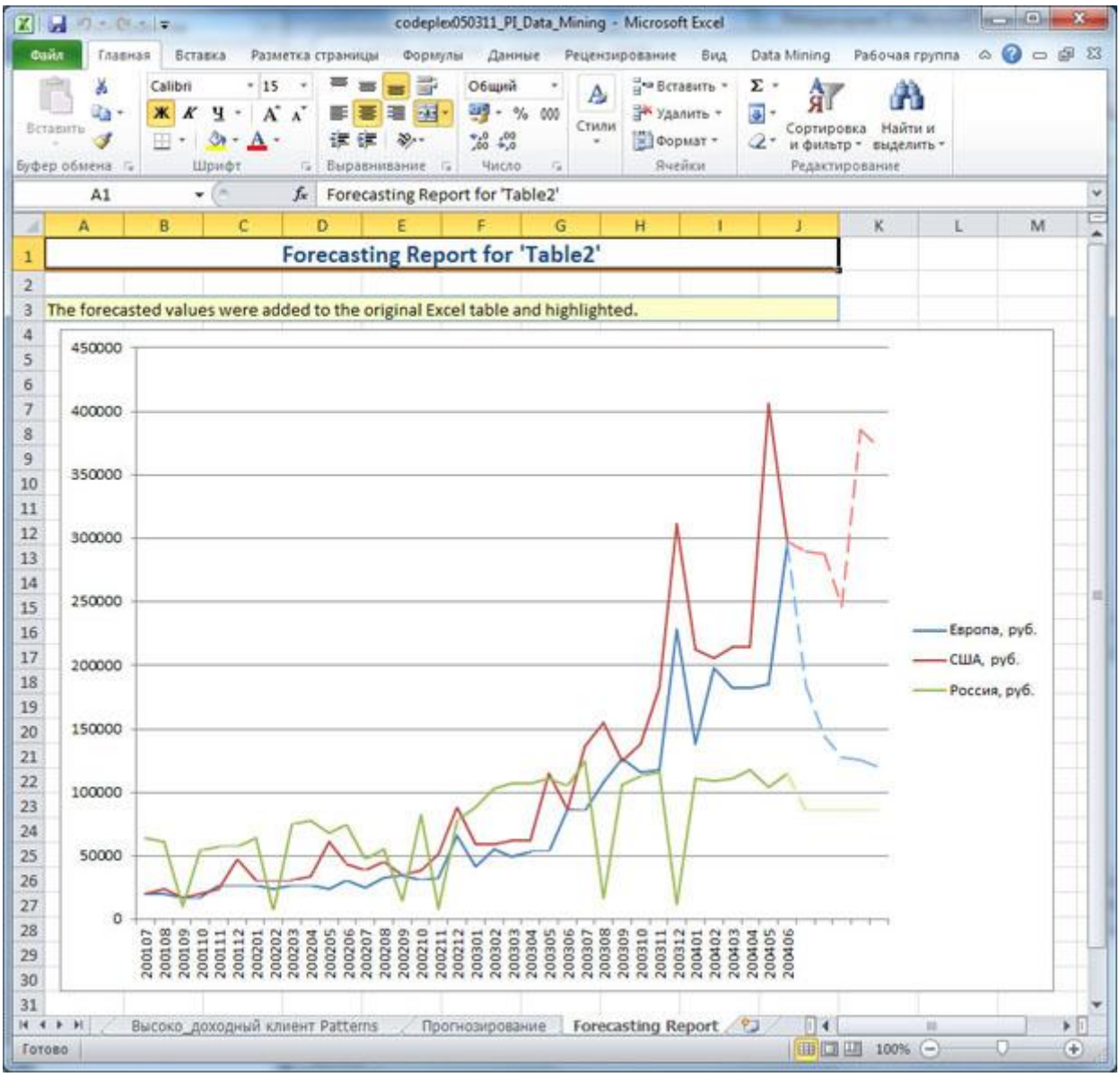


Рис. 6.7. Отчет инструмента "Прогноз"

Кроме того, в исходную таблицу добавляются результаты прогноза (столько значений, сколько было указано при запуске - рис. 6.6). На рис. 6.8 они выделены светло-желтым фоном. Чтобы продолжить ряд временных меток, можно выделить несколько последних значений столбца "Год/месяц" и незаполненную область в строках с прогнозом, выбрать на панели управления в ленте "Главная" кнопку "Заполнить" (рис. 6.8 подчёркнута красным), из выпадающего списка выбрать вариант "Прогрессия" и указать автоматическое определение шага. Недостающие значения будут добавлены. Теперь на графике будут автоматически проставлены недостающие временные метки.

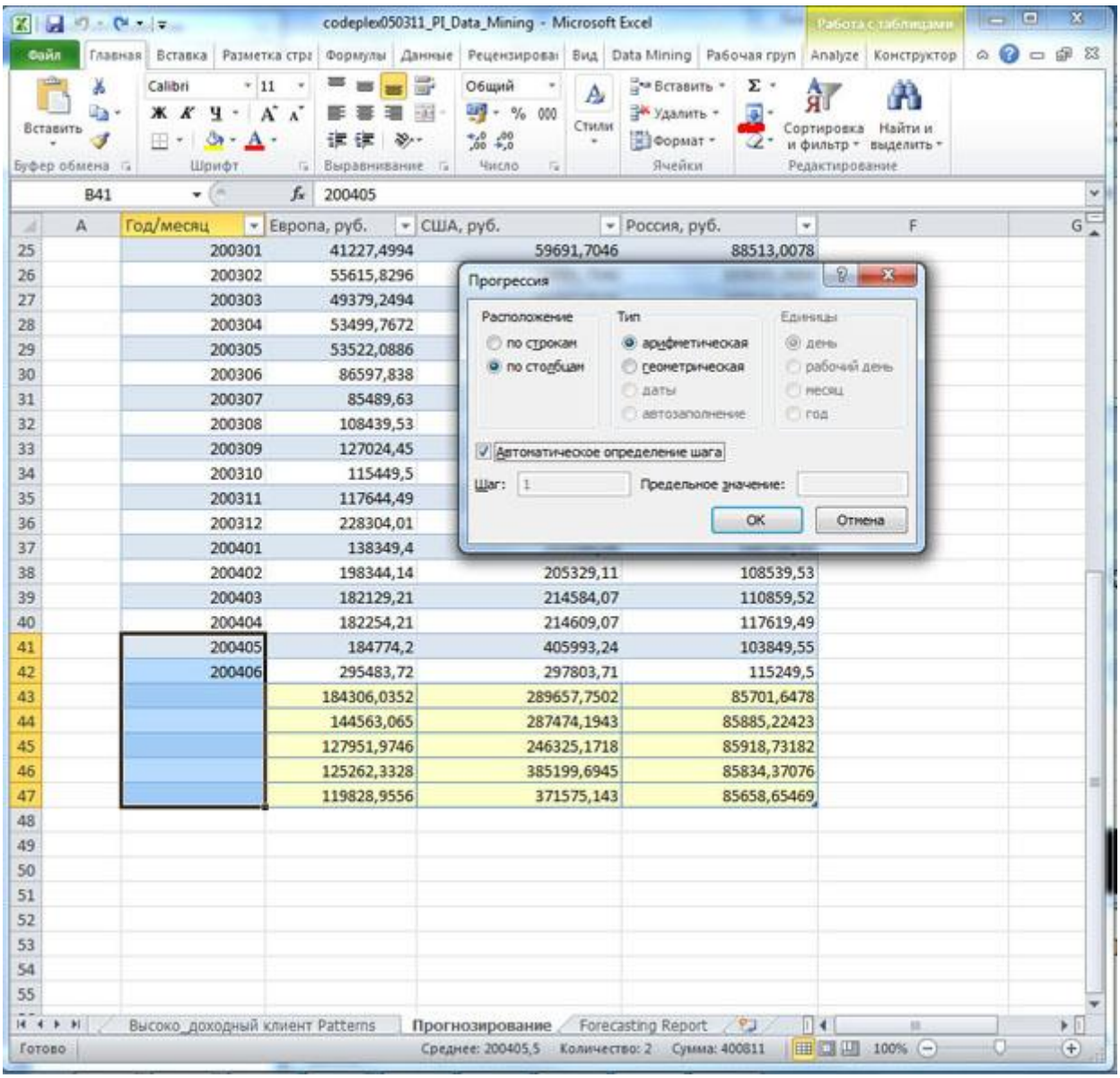


Рис. 6.8. Предсказанные значения и заполнение столбца временных меток

Чтобы убрать результаты работы инструмента, надо удалить лист отчета и строки исходной таблицы с предсказанными значениями.

Задание. С помощью инструмента постройте прогноз продаж на год (12 значений). Проанализируйте график. На ваш взгляд, какой тип периодичности обнаружил инструмент в исходных данных и использует для предсказания?

Лабораторная работа 4. Использование инструментов "HighlightExceptions" и "ScenarioAnalysis"

Цель: Лабораторная работа посвящена использованию инструментов "Выделение исключений" ("HighlightExceptions") и "Анализ сценариев" ("ScenarioAnalysis").

В качестве учебного набора данных, как и в прошлых лабораторных, будем использовать локализованный пример для Excel, взятый с http://russiandmaddins.codeplex.com/

Выделение исключений

Как следует из названия, инструмент позволяет выявить данные, выделяющиеся среди имеющегося набора. Это может быть полезно в ряде случаев. Во-первых, это могут быть ошибочные данные (например, результаты ошибки оператора при вводе каких-то значений). Во-вторых, исключения могут представлять отдельный интерес (как, например, в случае обнаружения мошеннических действий с банковскими картами и т.п.). Кроме того, анализ исключений может рассматриваться как предварительная часть интеллектуального анализа данных с помощью других методов. В частности, это позволяет исключить попадание нетипичных примеров в обучающую выборку.

В ходе работы инструмент HighlightExceptions создает временную модель интеллектуального анализа с использованием алгоритма MicrosoftClustering. Для каждой анализируемой строки оценивается степень принадлежности выявленным кластерам.Значения, находящиеся далеко от всех кластеров, помечаются как исключения.

При запуске инструмента можно отметить столбцы, не учитываемые при анализе. В рекомендациях по использованию [1,3] указывается, что желательно исключить из анализа столбцы с уникальными значениями (имена, идентификаторы), а также содержащие много пустых значений или произвольный текст. На рис. 7.1 видно, что при анализе набора данных "Клиенты" инструмент предлагает исключить из рассмотрения поле ID.

По итогам работы (а работает этот инструмент несколько дольше рассмотренных нами ранее) формируется отчет (рис. 7.2) и в исходном наборе данных исключения выделяются цветом (рис. 7.3).

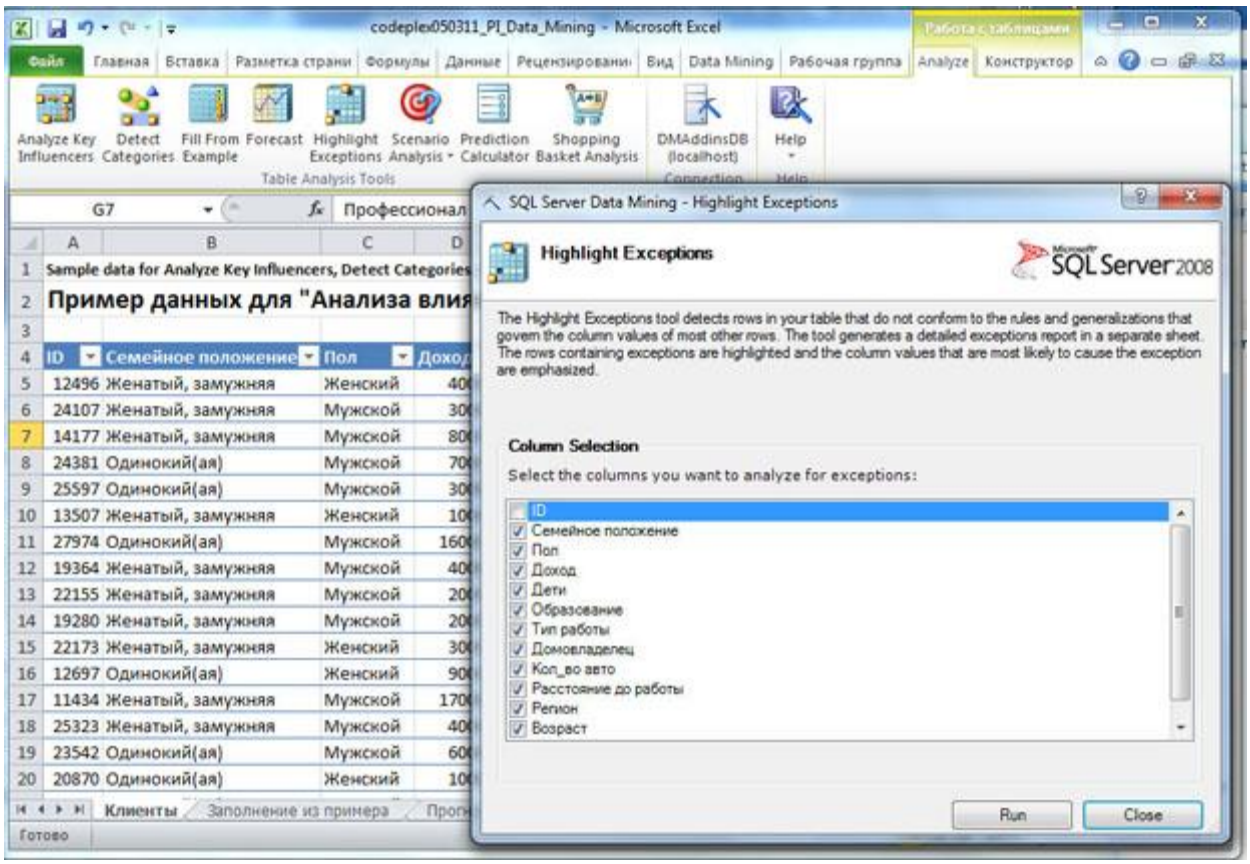


Рис. 7.1. Запуск инструмента HighlightExceptions

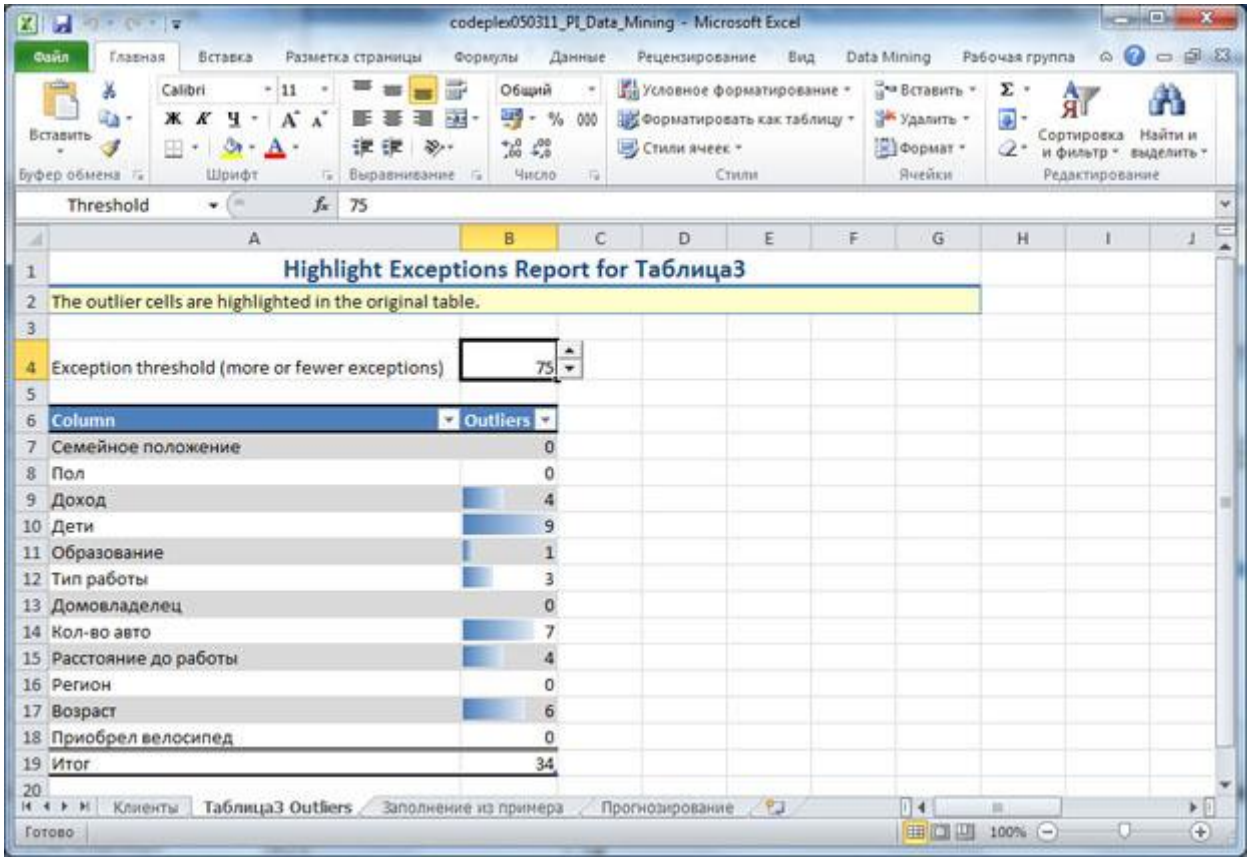


Рис. 7.2. Отчет по проведенному анализу данных

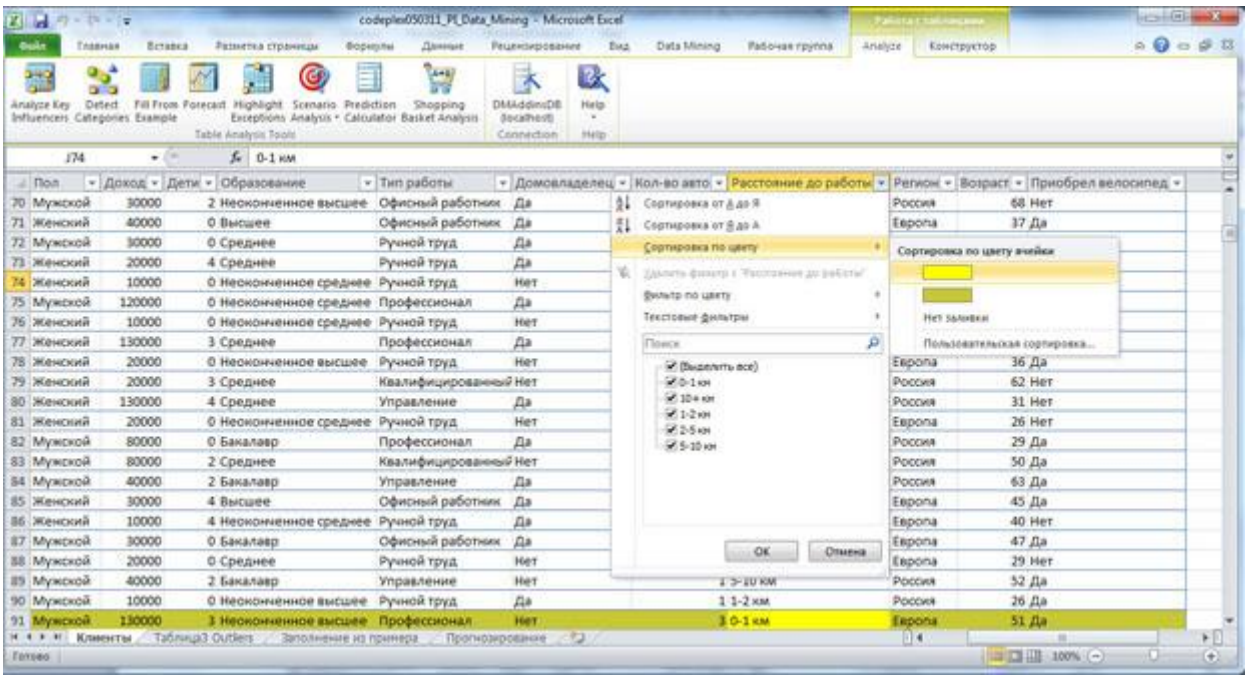


Рис. 7.3. Исключения выделяются цветом, что позволяет произвести сортировку

На рис. 7.2 видно, что инструмент позволяет указать порог отклонения от нормы (Exception threshold), измеряемый в процентах (оценка вероятности того, что выделенное значение относится к исключениям). Уменьшение порога приведет к тому, что больше записей будет рассматриваться как исключения, увеличение - наоборот. При значении по умолчанию в 75 % нашем наборе данных обнаружено 34 исключения. Отчет показывает, в каких столбцах сколько исключений было обнаружено.

Перейдем на лист Excel с данными. Рассматриваемые как выбросы значения выделяются в таблице цветом: вся строка-коричневым, конкретное значение - желтым. Чтобы сгруппировать нужные строки можно воспользоваться функциями Excel, позволяющими провести сортировку по цвету.

Также можно воспользоваться инструментами вкладки "Вид", чтобы создать новое окно и расположить рядом с окном с отчетом и данными (рис. 7.4). Пусть в отобранном наборе записей мы обнаружили ошибку. Скажем расстояние до работы у некоего клиента из США, обладающего двумя машинами, не "0-1 км", а "5-10 км" (именно поэтому ему нужно в семье 2 машины). Если мы изменим значение, будет произведен автоматический пересчет. В случае, представленном на рис. 7.4, новое значение уже не рассматривается как выброс.

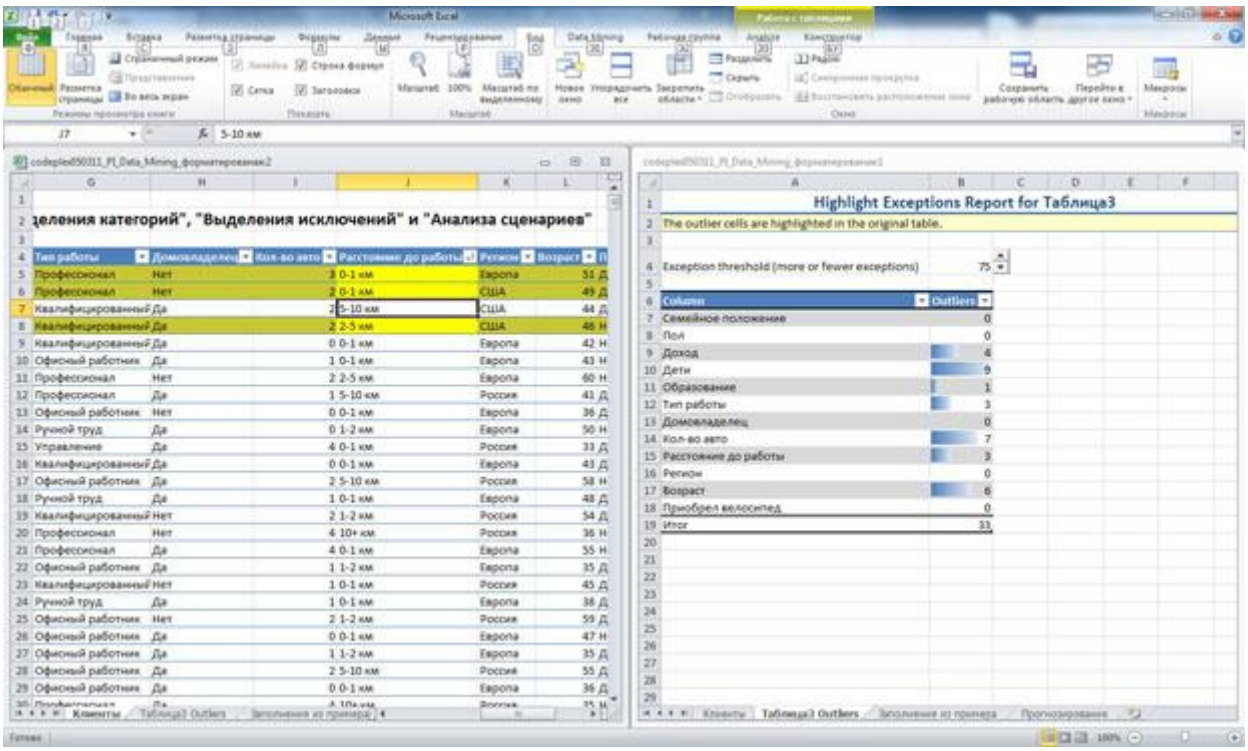


Рис. 7.4. Исправление ошибочного значения

Обратите внимание, что не только изменилась раскраска строки таблицы, но и произошли изменения в отчете, показывающем теперь наличие 33 исключений. Автоматический пересчет работает только в том случае, если сессия работы с аналитическими службами SQLServer остается открытой. Если таблица Excel была закрыта и снова открыта, то автоматического пересчета не будет (нужно снова провести анализ).

Также в описаниях отмечается, что инструмент реагирует только на изменения данных в диапазоне ячеек, использовавшемся при обучении. Если после начала работы инструмента в конец таблицы добавить новые строки, они оцениваться не будут.

Как уже отмечалось выше, если нужно рассматривать только наиболее сильные выбросы, можно увеличить значение порога отклонения и инструмент изменит оценки в соответствии с заданным значением (рис. 7.5).

Повторный запуск инструмента удалит результаты предыдущего анализа. Учитывая, что проводимые инструментом изменения достаточно сложны (раскраска строк таблицы и т.д.), если нужно удалить результаты работы, рекомендуется запустить повторный анализ, согласиться с удалением результатов и потом в окне, аналогичном представленному на рис. 7.1, нажать кнопку Close (отказаться от анализа данных).

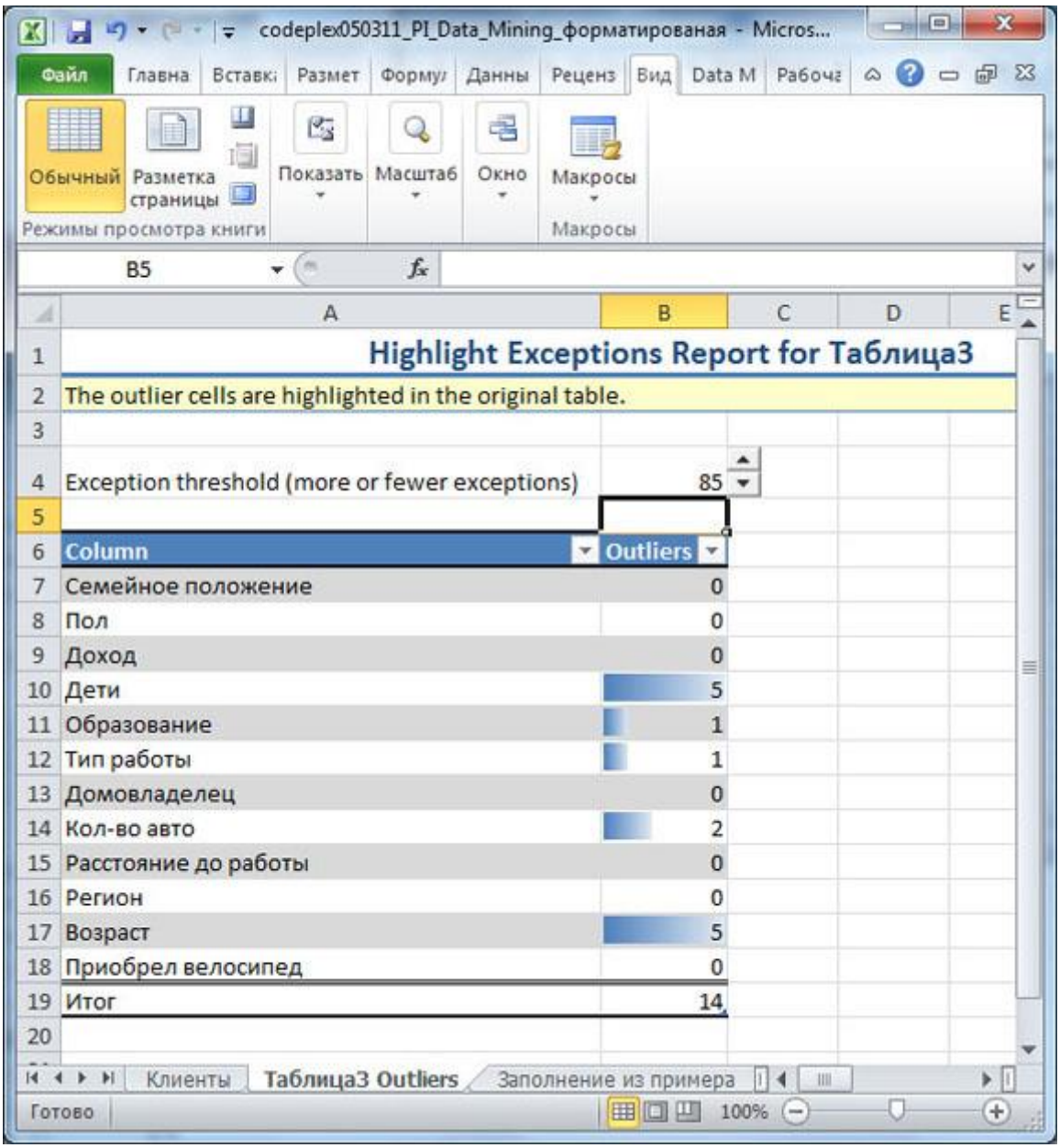


Рис. 7.5. Увеличение порога отклонения уменьшает число исключений

Задание. Проведите анализ исключений для набора данных "Клиенты" и значения порога в 90%. Предложите интерпретацию полученных результатов.

Задание 2. Проведите анализ исключений для набора данных "Прогнозирование" (продажи по месяцам в разных регионах). Предложите интерпретацию полученных результатов.

Анализ сценариев

Инструмент Scenario Analysis позволяет моделировать влияние, оказываемое изменением одного из параметров (значений одного столбца) на другой, связанный с первым. В основе работы инструмента лежит использование алгоритма Microsoft Logistic Regression. Для формирования временной модели требуется обучающая выборка, содержащая не менее 50 записей [3].

Инструмент Scenario Analysis включает две составные части - "Анализ сценария поиска решений" (GoalSeek) и "Анализ возможных вариантов"("What-If").

(i) "Анализ сценария поиска решений" (GoalSeek)

Использование инструмента GoalSeek позволяет оценить, сможем ли мы достичь желаемого значения в целевом столбце, меняя значения выбранного параметра. Инструмент позволяет провести анализ как для одной записи, так и для всей таблицы.

Используя этот инструмент надо быть готовым, что не для всех вариантов запроса может быть получен ответ. Это может быть связано с тем, что в исходных данных нет интересующих нас сочетаний. Также могут быть проблемы из-за типов данных.

Кроме того, нельзя забывать, что запрос нужно формировать с учетом знаний о предметной области. Например, можно запросить систему, если человек хочет увеличить годовой доход на 20 процентов, надо ли ему приобретать велосипед. И даже получить какой-то ответ. Но понятно, что в такой постановке сам вопрос является бессмысленным.

Пусть мы хотим узнать, как будет влиять образование на уровень достатка человека. Сначала проведем анализ для одной записи. Например, нас интересует клиент с идентификатором 12496 (первая запись в наборе данных).Откройте набор данных "Клиенты" и на вкладке Analysis выберите ScenarioAnalysis->GoalSeek (рис. 7.6).

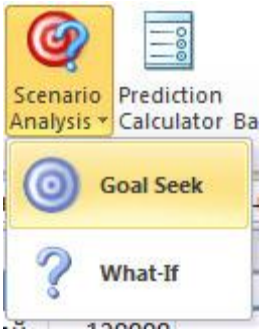


Рис. 7.6. Выбор инструмента GoalSeek

В окне параметров (рис. 7.7) укажем целевой столбец - "Доход", а также желаемое значение - 125% от текущего. В этом случае инструмент считает успешным результат, который не меньше заданного (в нашем примере $40000 \times 1,25 = 50000$ и более). Если задаваемое значение меньше 100%, то успешным считается результат, который не больше заданного. Также можно указать точное значение и диапазон (выбрав "Inrange"). Для значений, не являющихся числовыми, варианты "Percentage" и "Inrange" будут неактивны. Для достижения искомого значения будем менять столбец "Образование".

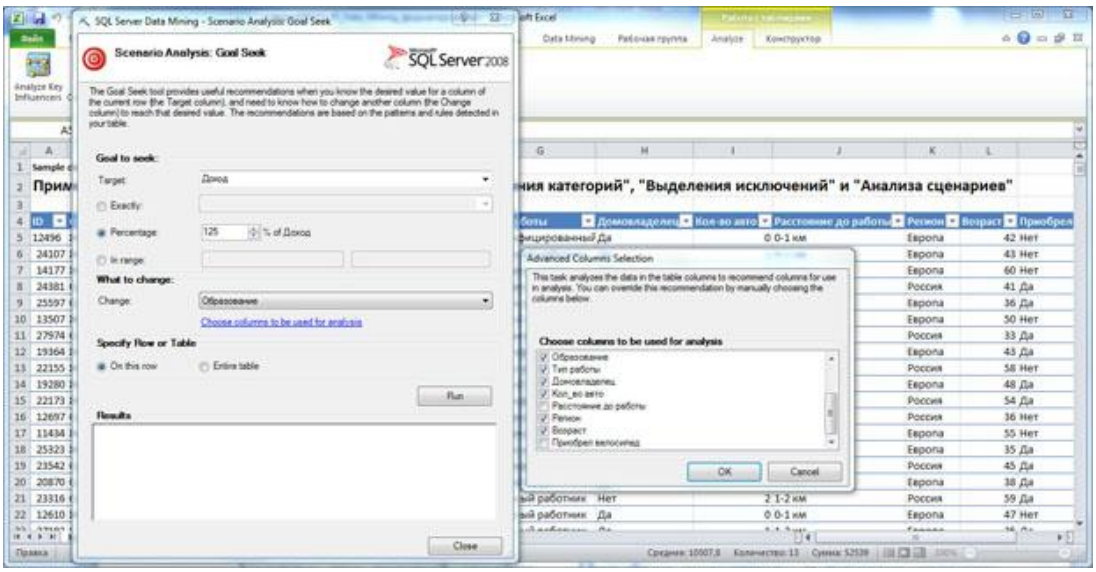


Рис. 7.7. Настройка параметров для GoalSeek

Перейдя по ссылке "Choose columns to be used for analysis", отметим, что при анализе в рассмотрение не берем столбцы "ID", "Дети", "Расстояние до работы", "Приобрел велосипед".После закрытия окна "Advanced Columns Selection" стоит еще раз проверить настройки в секции "Goaltoseek" - иногда при переходе между окнами переключатель между "Exactly", "Percentage" и "Inrange" сбрасываетсяв значение по умолчанию ("Exactly")

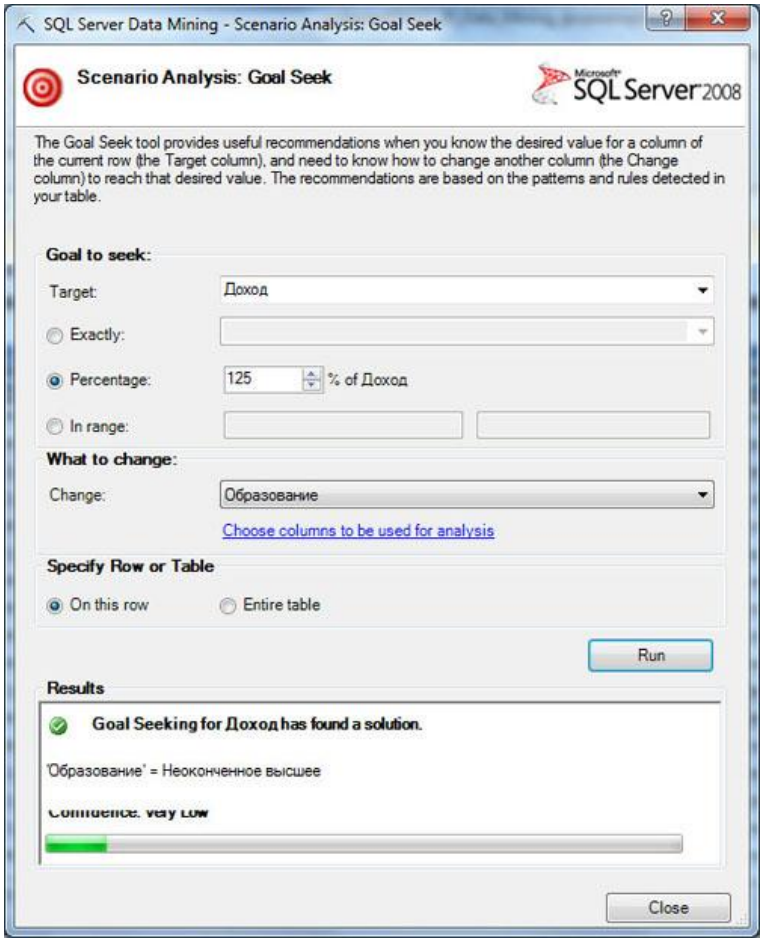


Рис. 7.8. Результат анализа для одной строки - решение найдено

Результат анализа, выполненного по нажатию кнопки Run, представлен на рис. 7.8. Для выбранной строки найден шаблон, рекомендующий для параметра "Образование" значение "Неоконченное высшее". При этом уровень достоверности - Confidence (иногда верхняя часть надписи затирается, как на рисунке), оценивается как очень низкий ("Very low").

Если прейти на следующую строку и снова нажать Run, получим результат для новых данных (рис. 7.9). В этом случае, подходящего решения не было найдено, и был предложен наиболее близкий вариант.

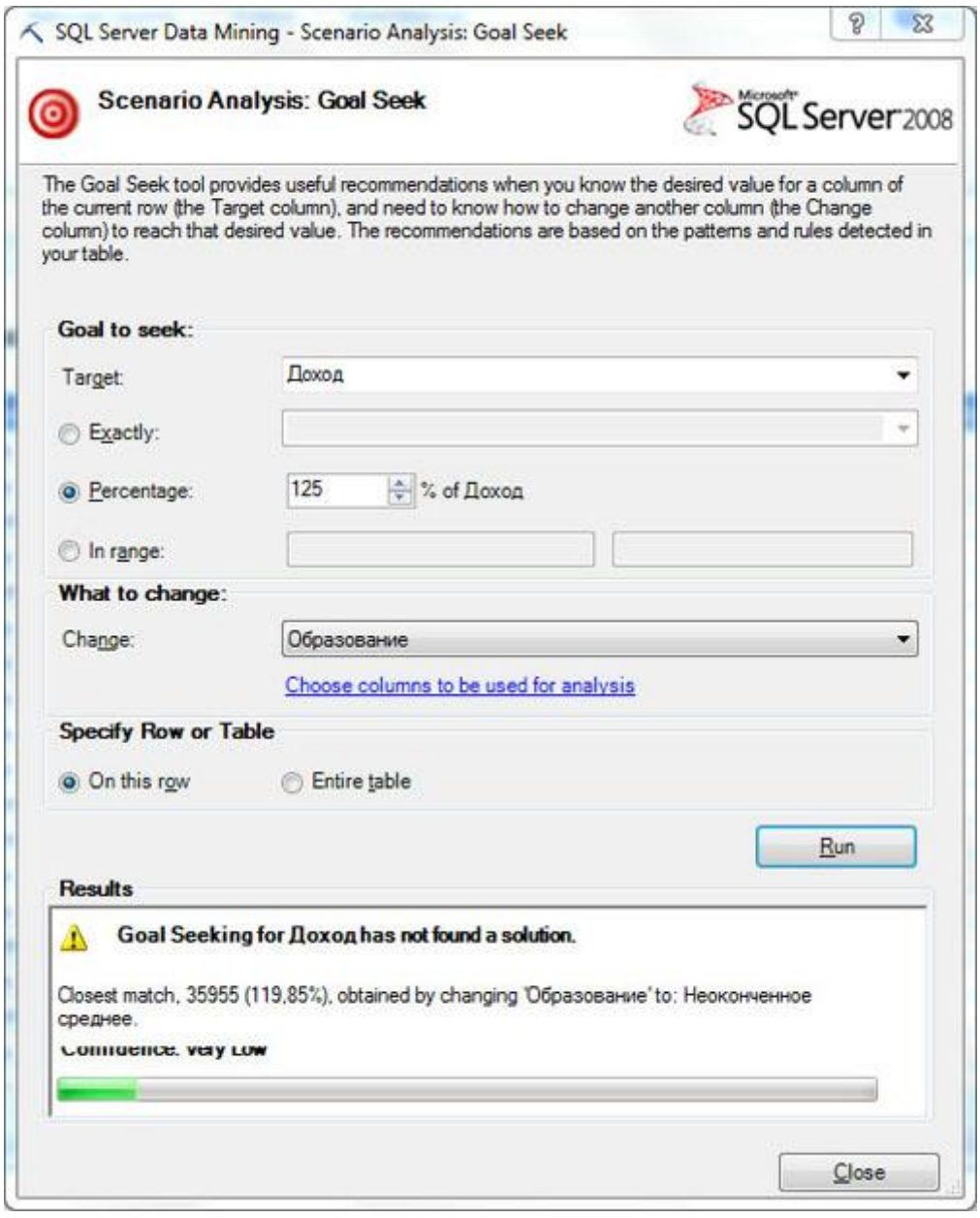
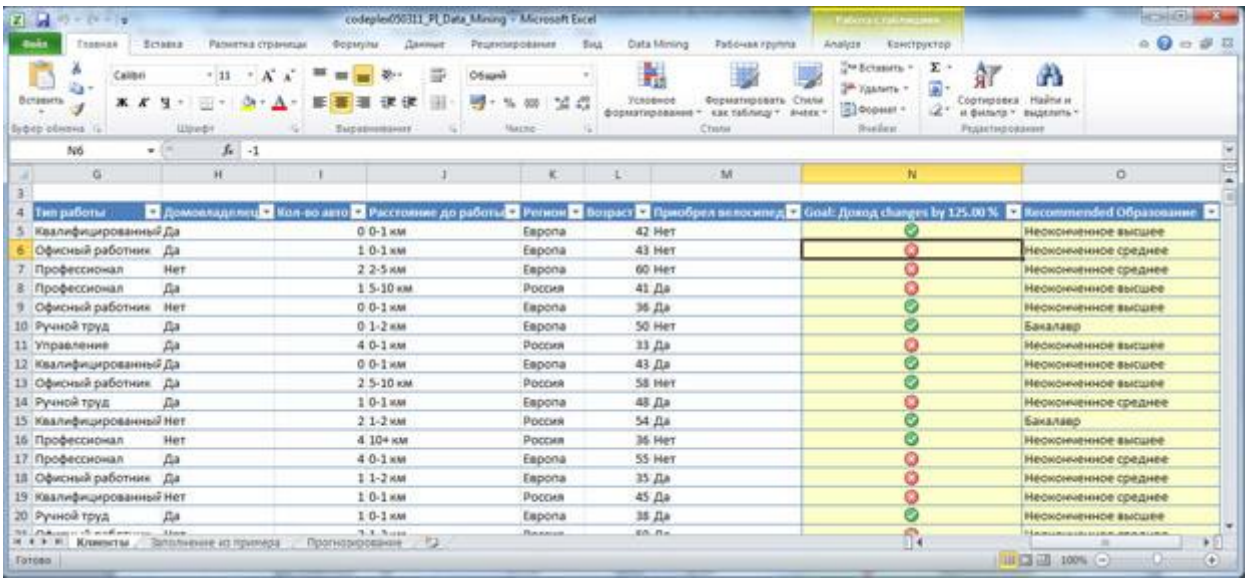


Рис. 7.9. Результат анализа для одной строки - решение не найдено



Тип работы	Домовладелец	Кол-во авто	Расстояние до работы	Регион	Возраст	Приобрел велосипед	Goal: Доход changes by 125.00 %	Recommended Образование
Квалифицированный	Да	0	0-1 км	Европа	42	Нет	✓	Неоконченное высшее
Офисный работник	Да	1	0-1 км	Европа	43	Нет	✗	Неоконченное среднее
Профессионал	Нет	2	2-5 км	Европа	60	Нет	✗	Неоконченное среднее
Профессионал	Да	1	5-10 км	Россия	41	Да	✗	Неоконченное высшее
Офисный работник	Нет	0	0-1 км	Европа	36	Да	✓	Неоконченное высшее
Ручной труд	Да	0	1-2 км	Европа	50	Нет	✓	Бакалавр
Управление	Да	4	0-1 км	Россия	33	Да	✗	Неоконченное высшее
Квалифицированный	Да	0	0-1 км	Европа	43	Да	✓	Неоконченное высшее
Офисный работник	Да	2	5-10 км	Россия	58	Нет	✓	Неоконченное высшее
Ручной труд	Да	1	0-1 км	Европа	48	Да	✗	Неоконченное среднее
Квалифицированный	Нет	2	1-2 км	Россия	54	Да	✓	Бакалавр
Профессионал	Нет	4	10+ км	Россия	36	Нет	✓	Неоконченное высшее
Профессионал	Да	4	0-1 км	Европа	55	Нет	✗	Неоконченное среднее
Офисный работник	Да	1	1-2 км	Европа	35	Да	✗	Неоконченное среднее
Квалифицированный	Нет	1	0-1 км	Россия	45	Да	✗	Неоконченное среднее
Ручной труд	Да	1	0-1 км	Европа	38	Да	✓	Неоконченное высшее

Рис. 7.10. Анализ для всей таблицы

А если в секции "Specify Rowor Table" установить переключатель в "Entire table", то сценарии будут посчитаны для всех строк (рис. 7.10). Результаты будут указаны в двух столбцах, добавленных в исходную таблицу. Для тех строк, которые отмечены крестиком в красном круге, соответствующего желаемому сценарию шаблона найдено не было.

Задание. Проведите анализ для отдельной строки и таблицы, аналогичный описанному выше. Прокомментируйте результаты.

Примечание. Запуск процедуры анализа для ряда других комбинаций столбцов (например - целевой столбец "покупка велосипеда" = "да", независимая переменная - "расстояние до работы") приводит к ошибке "Query (1, 50) Синтаксический анализатор: Неверный синтаксис "value".", видимо связанной с некорректной обработкой некоторых типов данных.

(ii) "Анализ возможных вариантов"("What-If")

Инструмент What-If позволяет решить обратную по отношению к GoalSeek задачу: оценить значение целевой переменной при определенном изменении заданного параметра.

Например, можно оценить, как изменился бы уровень дохода человека, если бы повысился его уровень образования. Перейдем на запись с идентификатором 12697 и запустим инструмент: Scenario Analysis->What-If. Укажем параметры сценария: образование меняется на "Высшее" и целевой столбец "Доход". Полученный для строки результат показывает, что при изменении уровня образования доход может несколько вырасти (исходное значение 90000, среднее значение для нового шаблона 104448). Но степень уверенности в прогнозе не слишком высокая.

Аналогично предыдущему инструменту, подобный анализ сценария можно сделать и для всей таблицы целиком. В этом случае к исходной таблице добавляются два столбца - один показывает новое значение целевого параметра, второй - оценку достоверности (рис. 7.12). Достоверность

оценивается числом от 0 до 100: 100 - максимальная достоверность (абсолютная уверенность в прогнозе), 0 - минимальная.

Задание. Проведите анализ данных, аналогичный описанному выше.

Для того чтобы удалить результаты работы с таблицей инструментов What-If и Scenario Analysis, достаточно удалить добавленные столбцы. При работе с отдельными строками, никаких дополнительных действий не требуется.

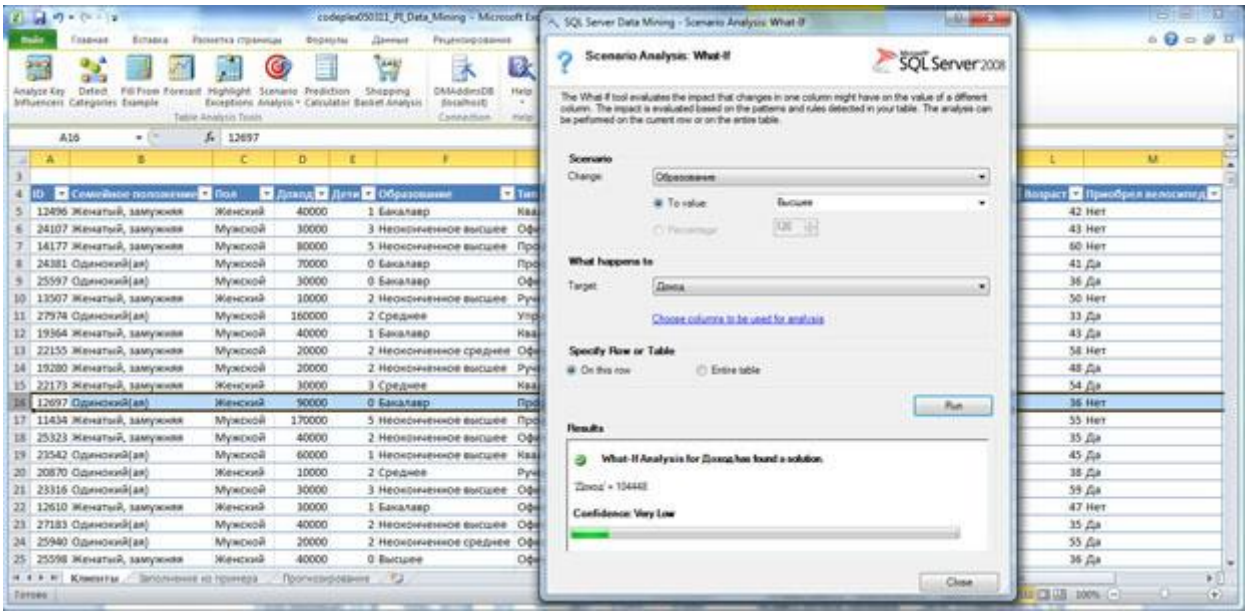


Рис. 7.11. Работа инструмента What-If для отдельной строки

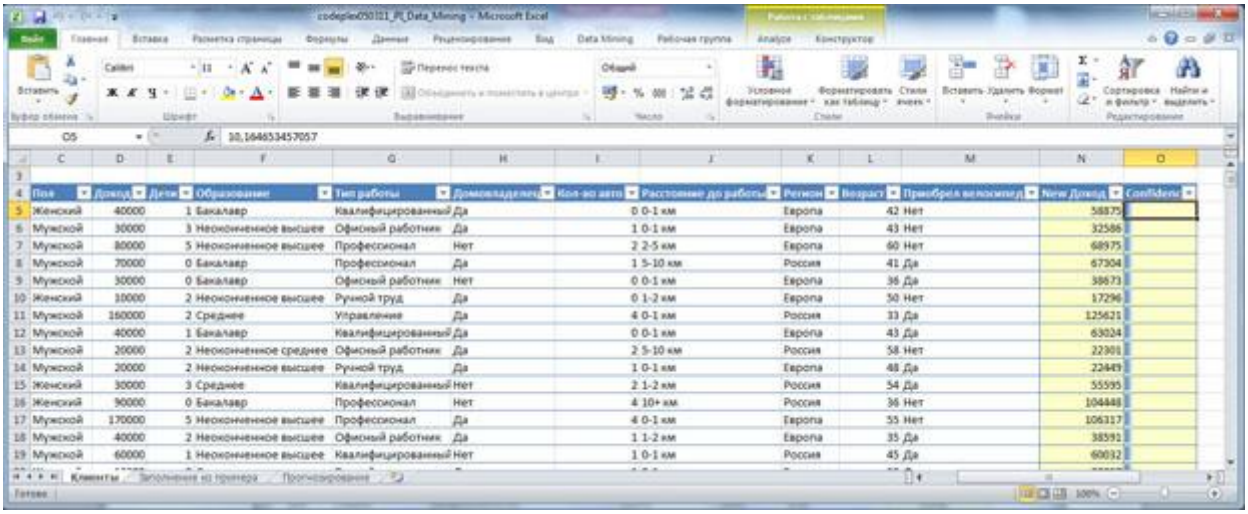


Рис. 7.12. Прогноз What-Ifдля всей таблицы

Лабораторная работа 5. Использование инструментов "Prediction Calculator" и "ShoppingbasketAnalysis"

Цель: Лабораторная работа посвящена использованию инструментов "Расчет прогноза" ("PredictionCalculator") и "Анализ покупательской корзины" ("ShoppingBasketAnalysis").

Расчет прогноза

Инструмент Prediction Calculator помогает сгенерировать и настроить "калькулятор", который позволяет оценить шансы на получение ожидаемого значения целевого параметра без подключения к аналитическим службам SQLServer. В частности, такая возможность может быть очень полезна для удаленных пользователей.

В качестве учебного набора данных в этой части лабораторной будем использовать локализованный пример для Excel, взятый с http://russiandmaddins.codeplex.com/

Перейдем на набор данных "Клиенты" и на вкладке Analyze выберем Prediction Calculator. В окне настроек надо указать целевой столбец и искомое значение (рис. 8.1). Если значения целевого столбца рассматриваются как числовые из непрерывного диапазона, то можно указать, как точное значение, так и желаемый интервал. В противном случае - только точное значение.

Рис. 8.1. Настройки инструмента Prediction Calculator

Пусть цель анализа - определить, купит ли клиент велосипед. В качестве целевого столбца указываем "Приобрел велосипед" и значение "Да". Далее можно указать столбцы для анализа. Как и ранее, рекомендуется исключать из рассмотрения столбцы с уникальными значениями и столбцы, один из

которых дублирует другой (например, точное значение заработной платы и диапазон заработной платы).

Инструмент всегда формирует отчет Prediction Calculator Report, кроме того по умолчанию формируются два необязательных отчета - Prediction Calculator ("калькулятор" прогноза в виде таблицы Excel) и Printable Calculator (таблица калькулятора для печати и ручной обработки).

Чтобы лучше разобраться с результатами работы инструмента, перейдем сначала на лист с отчетом Prediction Calculator.В верхней части отчета расположен сам калькулятор (рис. 8.2), в нижней - таблица баллов, соответствующих различным значениям параметров (рис. 8.3).

Работая с калькулятором, можно описать анализируемый пример, указывая значения для каждого параметра. Значения в столбец Value можно вводить или выбирать из выпадающих списков (что лучше, т.к. меньше шансов ввести некорректное значение или диапазон). Для описываемого примера рассчитывается сумма баллов, которая сравнивается с рекомендуемым пороговым значением. Если значение выше "порога", то прогноз получает значение "истина" (на рисунке сумма баллов 572, пороговое значение 565). Вторая часть отчета поясняет полученный результат, показывая, сколько баллов за какое значение ставится.

Prediction Calculator for the 'Да' state of 'Приобрел велосипед'

Suggested Threshold to maximize profit:		565

Attribute	Value	Relative Impact
Семейное положение	Женатый, замужняя	0
Пол	Мужской	0
Доход	39050 - 71062	58
Дети	0	168
Образование	Бакалавр	19
Тип работы	Профессионал	114
Домовладелец	Да	25
Кол_во авто	Да / Нет	90
Расстояние до работы	0-1 км	83
Регион	США	0
Возраст	< 37	15
Итог		572

Prediction for 'Да'	ИСТИНА

Рис. 8.2. "Калькулятор"

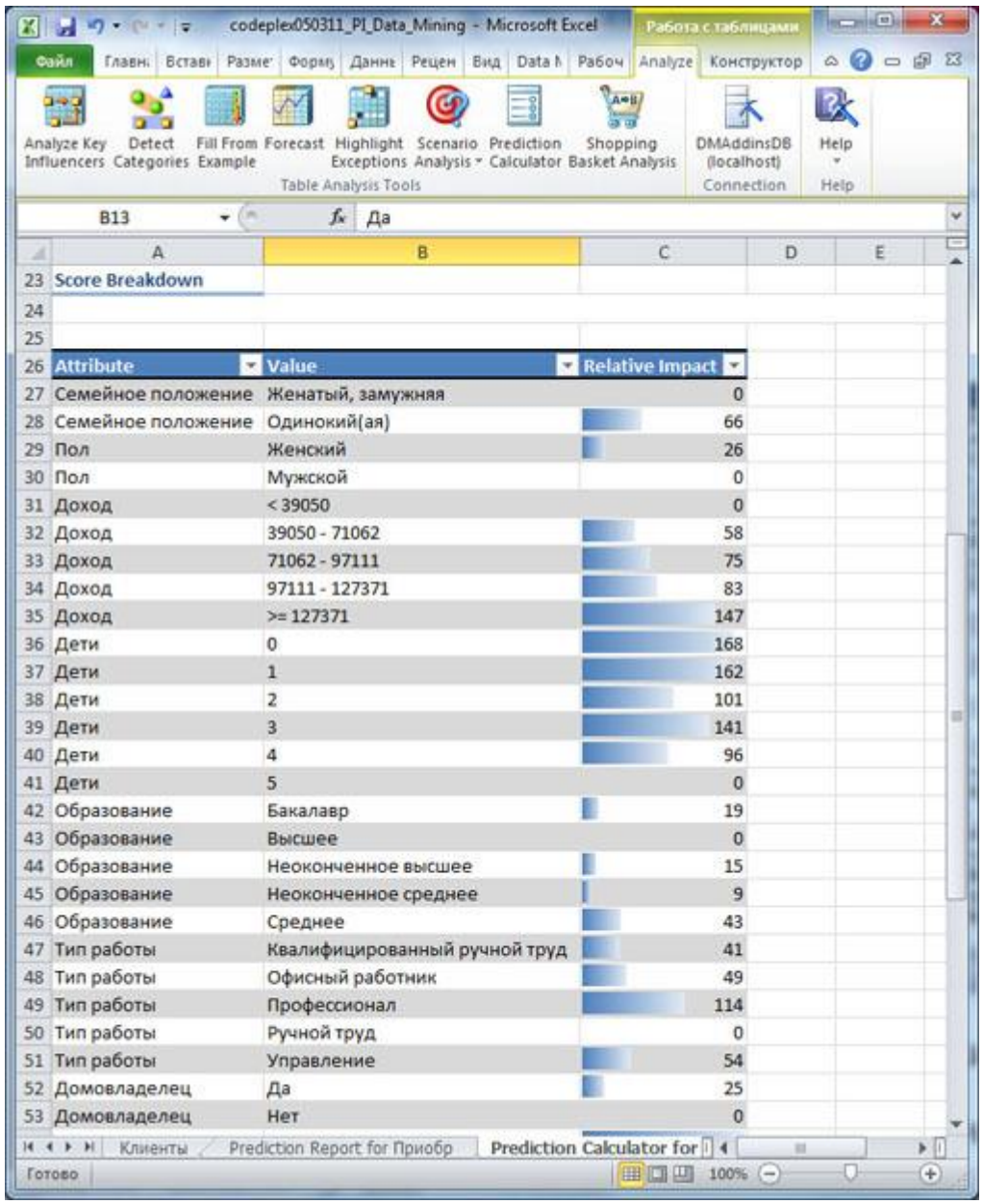


Рис. 8.3. Таблица баллов для параметров

Представленный на рис. 8.4 отчет "Printable Calculator" позволяет вывести на печать готовую форму для ручного подсчета баллов и получения оценки без использования компьютера. Это может быть удобно, например, для торговых представителей или других сотрудников, работающих вне офиса и не имеющих доступа к компьютеру. Все что нужно для расчета прогноза - отметить варианты, просуммировать баллы и сравнить с пороговым значением.

Теперь перейдем к более интересному вопросу - как же было определено пороговое значение. Отчет Prediction Calculator Report позволяет с этим разобраться (рис. 8.5). По итогам анализа формируется прогноз, который может быть отнесен к одной из четырех категорий [1]:

- истинный позитивный прогноз (TruePositive) - верный прогноз. Например, клиент, для которого прогноз показал истину, на самом деле заинтересован в покупке велосипеда. Магазин получил прибыль;

- истинный негативный прогноз (TrueNegative) - верный негативный прогноз. Клиент, для которого прогноз показал незаинтересованность в покупке, на самом деле не собирается покупать велосипед. Магазин не получил прибыли, но и не понес затрат (на рассылку рекламных предложений и проч.);
- ложный позитивный прогноз (FalsePositive; ошибка 1 рода) - неверный прогноз, показывающий, что клиент хочет сделать покупку, хотя на самом деле это не так (может привести магазин к затратам на сопровождение клиента);
- ложный негативный прогноз (FalseNegative; ошибка 2 рода) - неверный прогноз, показывающий, что клиент не хочет сделать покупку, хотя на самом деле он в ней заинтересован (может привести к упущенной прибыли).

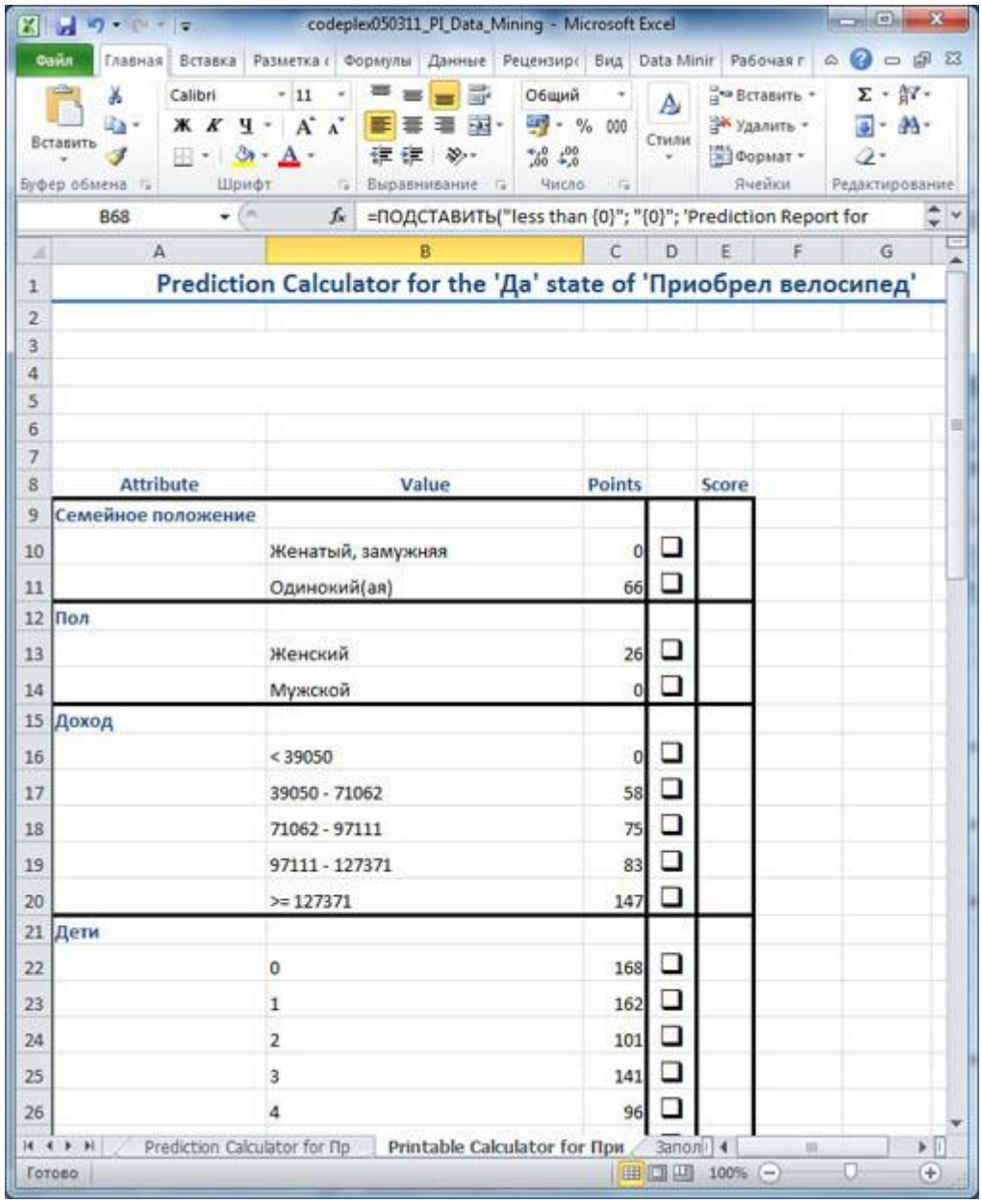


Рис. 8.4. Отчет "PrintableCalculator"

Отчет Prediction Calculator Report позволяет указать прибыль от истинных прогнозов и убыток от ложных. На основе этих данных определяется пороговое значение, обеспечивающее максимум прибыли. По умолчанию, для истинного позитивного прогноза указывается прибыль 10 (долларов или других единиц), для ложного позитивного - такой же убыток (рис. 8.5, таблица в левой верхней части экрана). В этом случае максимум прибыли (график на рис. 8.5 справа вверху) как раз и будет соответствовать пороговому значению для прогноза в 565 баллов.

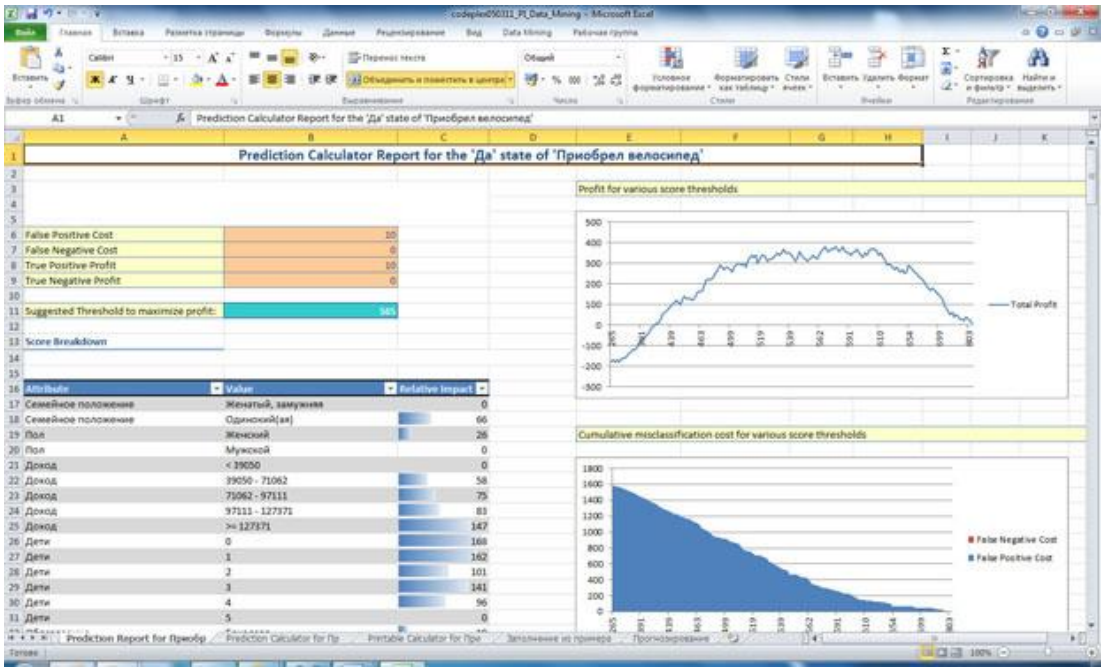


Рис. 8.5. Отчет Prediction Calculator Report

В нижней части отчета Prediction Calculator слева располагается таблица с относительными весами значений рассматриваемых параметров (ее мы уже встречали в таблице Prediction Calculator) и графиком потерь от ложных прогнозов.

Пусть продажа велосипеда приносит магазину не 10, а 50 долларов. В этом случае, прибыль от одной продажи будет перекрывать затраты на сопровождение до 5 отказавшихся от покупки клиентов. Соответственно изменится и соотношение прибыли/затраты. На рис. 8.6 показано, что в этом случае, для максимизации прибыли рекомендуется установить пороговое значение для прогноза в 443 балла. Новое значение будет автоматически подставлено и в таблицу Prediction Calculator.

Задание. Проведите анализ для двух различных наборов значений прибыли от истинных прогнозов и убытков от ложных. Прокомментируйте результаты.

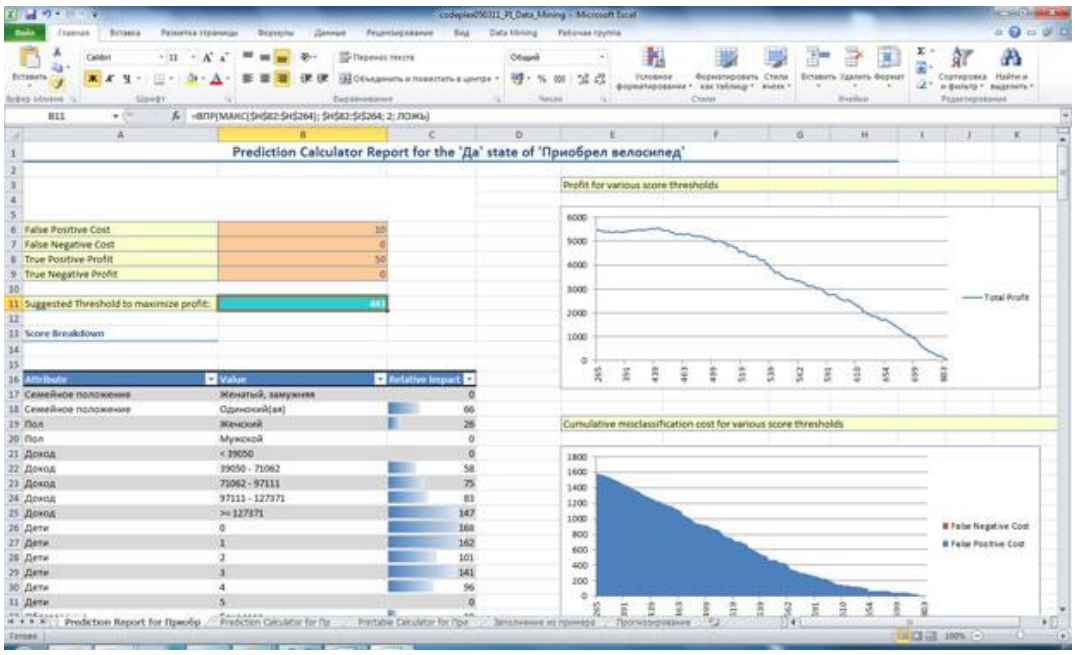


Рис. 8.6. Отчет Prediction Calculator Report: при вводе новой оценки прибыли от правильного прогноза меняется рекомендуемое пороговое значение

Анализ покупательской корзины

В наборе Table Analysis Tools нам осталось рассмотреть инструмент Shopping Basket Analysis. Он позволяет, например, на основе данных о покупках выделить товары, чаще всего встречающиеся в одном заказе, и сформировать рекомендации относительно совместных продаж.

В процессе анализа используется алгоритм MicrosoftAssociationRules.

Для изучения этого инструмента, вместо использованного ранее локализованного набора данных, обратимся к примеру, из поставки надстроек интеллектуального анализа (в предыдущем файле нужного набора данных просто нет). Через меню "Пуск" найдите "Надстройки интеллектуального анализа данных" -> "Образцы данных Excel". В этой книге Excel с первого листа (рис. 8.7) перейдите по ссылке "Поиск взаимосвязей и покупательское поведение". Соответствующий набор данных (рис. 8.8) содержит информацию о заказах (номер заказа - Order Number), включенных в них товарах (их категории - Category и собственно товаре - Product) и ценах.

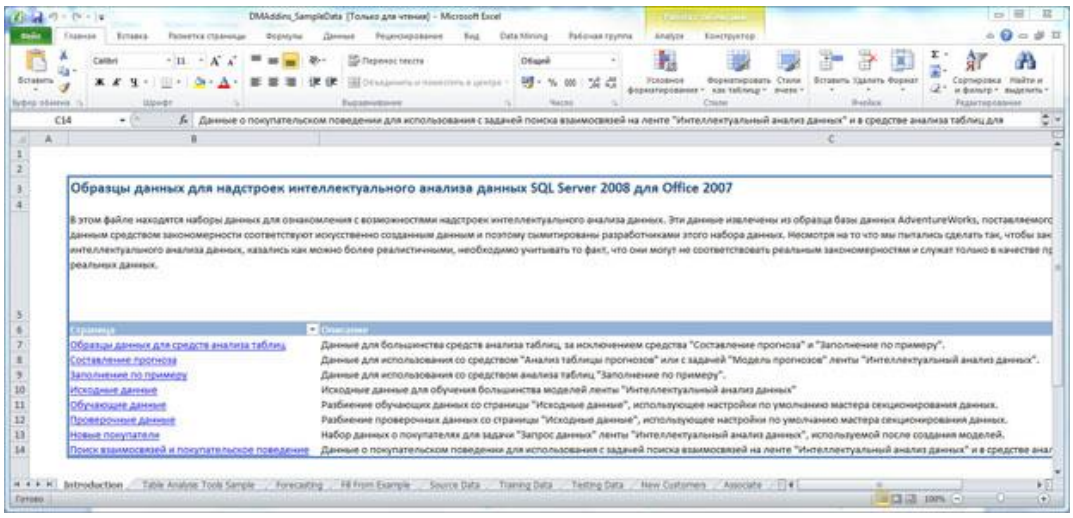


Рис. 8.7. Образцы данных

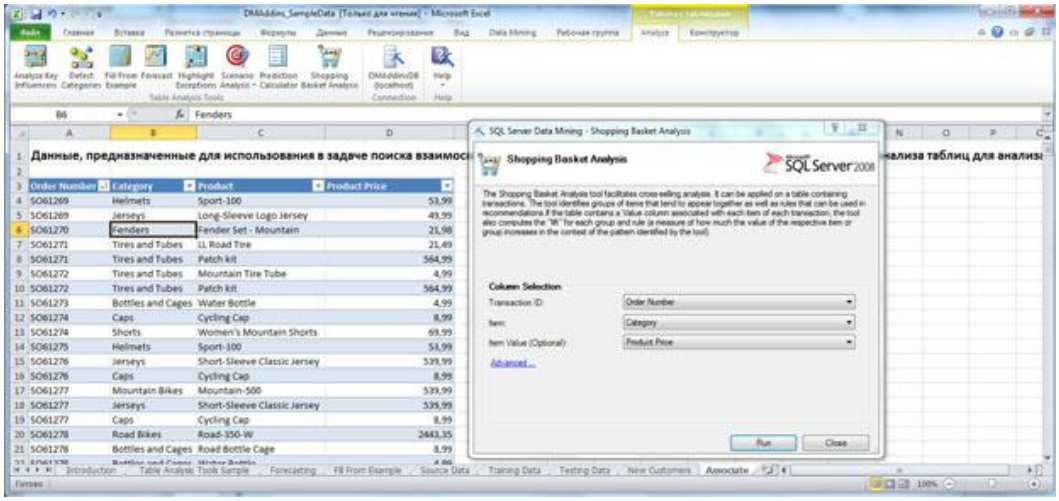
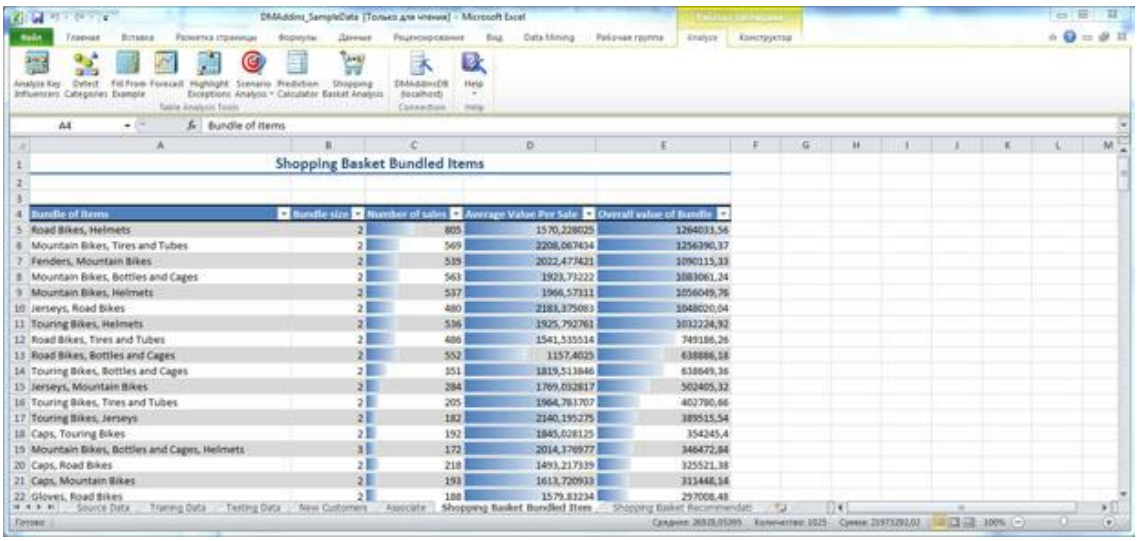


Рис. 8.8. Запуск инструмента Shopping Basket Analysis

Проанализируем, какие категории чаще всего попадают в один заказ. Запустим инструмент Shopping Basket Analysis. В его настройках надо указать идентификатор транзакции (TransactionID), в нашем случае, это Order Number и предмет анализа (мы будем проводить анализ для категорий - Category). Необязательным параметром, количественно характеризующим предмет анализа (Item Value), в нашем случае будет цена. Если Item Value не указан, то анализироваться будет только частота выявленных сочетаний.

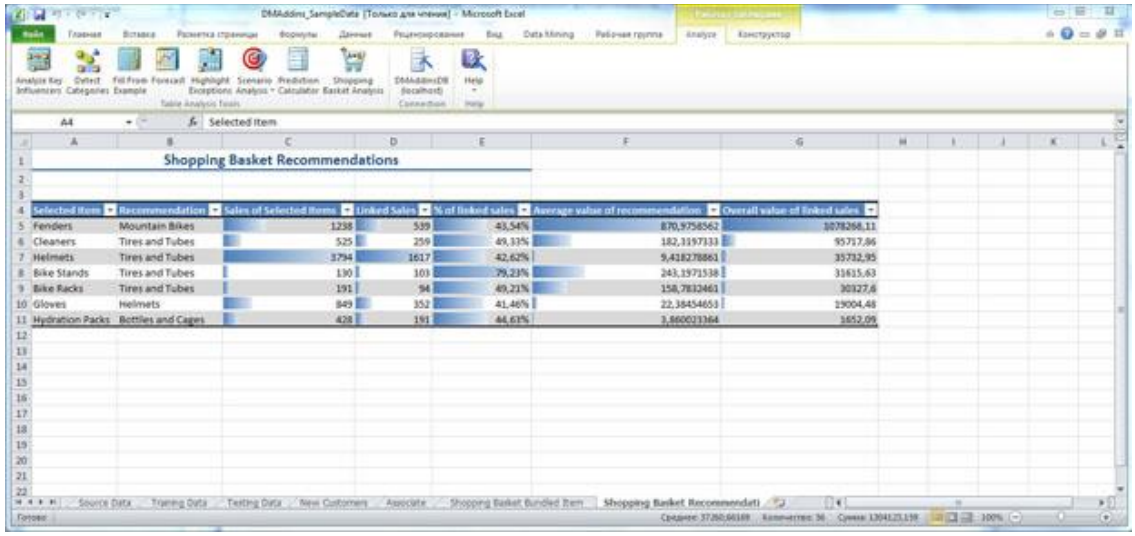
Результаты работы Shopping Basket Analysis отображаются в двух отчетах - Bundled Items и Recommendations.



Shopping Basket Bundled Items

Bundle of Items	Bundle size	Number of sales	Average Value Per Sale	Overall value of Bundle
Road Bikes, Helmets	2	805	1570,228025	1264033,54
Mountain Bikes, Tires and Tubes	2	569	2208,067434	1256390,37
Fenders, Mountain Bikes	2	539	2022,477421	1090115,33
Mountain Bikes, Bottles and Cages	2	563	1923,73222	1083061,24
Mountain Bikes, Helmets	2	537	1966,57311	1056049,76
Jerseys, Road Bikes	2	480	2183,375083	1048020,04
Touring Bikes, Helmets	2	536	1925,792761	1032224,92
Road Bikes, Tires and Tubes	2	486	1541,535514	749186,26
Road Bikes, Bottles and Cages	2	552	1157,4025	638886,18
Touring Bikes, Bottles and Cages	2	351	1819,511846	638649,36
Jerseys, Mountain Bikes	2	284	1769,032817	502405,32
Touring Bikes, Tires and Tubes	2	205	1964,781707	402780,66
Touring Bikes, Jerseys	2	182	2140,195275	389515,54
Caps, Touring Bikes	2	192	1845,028125	354245,4
Mountain Bikes, Bottles and Cages, Helmets	3	172	2014,176977	346472,84
Caps, Road Bikes	2	218	1493,217339	325521,38
Caps, Mountain Bikes	2	193	1613,720933	311448,14
Gloves, Road Bikes	2	188	1579,81234	297008,48

Рис. 8.9. Отчет Shopping Basket Bundled Items



Shopping Basket Recommendations

Selected Item	Recommendation	Sales of Selected Items	Linked Sales	% of linked sales	Average value of recommendation	Overall value of linked sales
Fenders	Mountain Bikes	1238	539	43,54%	870,9758562	1078268,11
Cleaners	Tires and Tubes	525	259	49,33%	182,3197133	95717,96
Helmets	Tires and Tubes	3794	1617	42,62%	9,418278861	35732,95
Bike Stands	Tires and Tubes	130	103	79,23%	243,1971538	31615,63
Bike Racks	Tires and Tubes	191	94	49,21%	158,7832461	30327,6
Gloves	Helmets	849	352	41,46%	22,38454653	19004,48
Hydration Packs	Bottles and Cages	428	191	44,63%	3,860023364	1652,09

Рис. 8.10. Отчет Shopping Basket Recommendations

Первый из отчетов содержит информацию о наиболее часто встречающихся в "одном чеке" сочетаниях категорий товаров. Так, например, в первой строке отчета на рис. 8.9 мы видим, что чаще всего встречается сочетание категорий "дорожные велосипеды" и "шлемы" (RoadBikes, Helmets). В рассматриваемом наборе оно выявлено в 805 заказах. Дальше указывается средняя цена набора и суммарная стоимость всех подобных наборов. Можно сказать, что этот отчет описывает покупательские шаблоны клиентов.

Второй отчет Shopping Basket Recommendations содержит рекомендации о товарах, которые могут быть предложены вместе. Например, третья строчка отчета указывает, что людям, купившим шлем, стоит также предложить приобрести шины. Это заключение базируется на том, что среди 3794 покупок, включающих шлемы, 1617 включали и шины. Доля таких связанных продаж равна 42,62%. Далее приводится средний доход от связанных продаж (общая стоимость, деленная на число транзакций, которые содержат "рекомендующий" продукт, в нашем случае - шлем) и общая сумма

связанных продаж. Основываясь на подобном отчете, владелец магазина может решить, как разместить товары, какие связанные предложения можно сформировать и т.д.

Для удаления результатов работы инструмента достаточно удалить сформированные отчеты.

Задание 1. Проведите анализ аналогичный описанному выше.

Задание 2. Проанализируйте, какие товары (а не категории товаров, как было раньше), приобретаются вместе. Опишите полученные результаты.

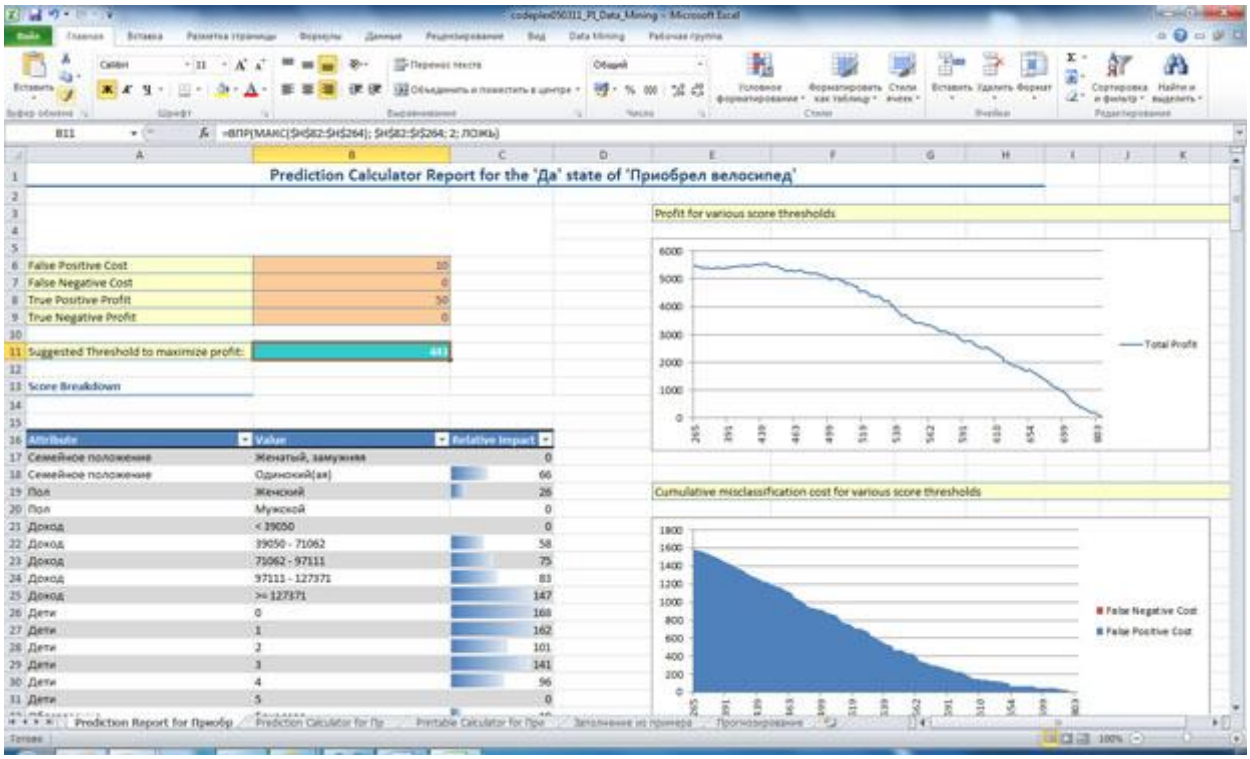


Рис. 8.6. Отчет Prediction Calculator Report: при вводе новой оценки прибыли от правильного прогноза меняется рекомендуемое пороговое значение

Лабораторная работа 6. Использование инструментов Data Mining Client для Excel для подготовки данных.

Цель: Данная лабораторная работа описывает возможности инструментов, относящихся к Data Mining Client для Excel 2007, в части подготовки данных для анализа.

Рассмотренные в предыдущих лабораторных работах "Средства анализа таблиц для Excel" (TableAnalysisTools) позволяют быстро провести "стандартный" анализ имеющихся данных. В то же время, этот набор инструментов не предоставляет особых возможностей по подготовке данных к анализу, оценке результатов и т.д. Из Excel это можно сделать, используя клиент интеллектуального анализа данных (DataMiningClient), который также входит в набор надстроек интеллектуального анализа. В ходе "Надстройки интеллектуального анализа данных для MicrosoftOffice" , отмечалось, что желательно сделать полную установку надстроек, в которую входит и DataMiningClient.

Откроем уже использовавшийся нами набор данных, входящий в поставку надстроек (меню "Пуск", найдите Надстройки интеллектуального анализа данных->Образцы данных Excel).Чтобы можно было спокойно вносить изменения, лучше сохранить его под новым именем.Перейдите на лист "Исходные данные" (SourceData) и щелкните на закладке DataMining. Лента с предлагаемыми инструментами представлена на рис. 13.1.

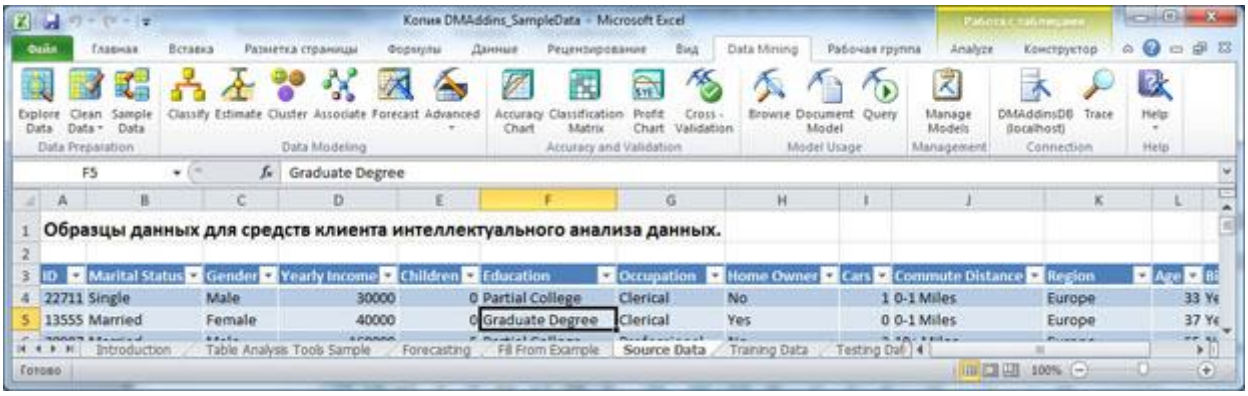


Рис. 13.1. Инструменты Data Mining Client

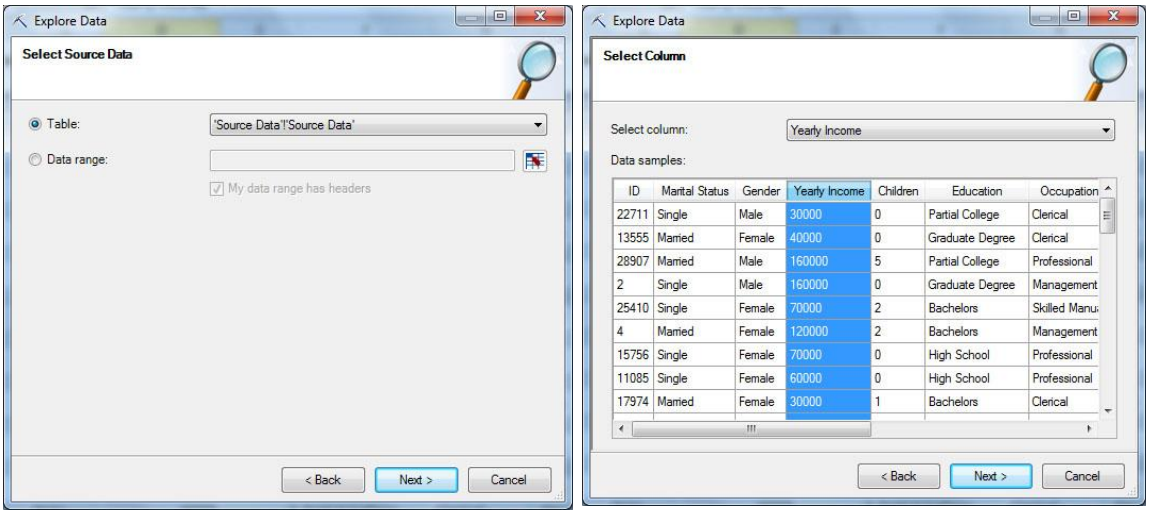
Первая группа инструментов (Data Preparation - Подготовка данных), позволяет провести первое знакомство с набором данных и подготовить его для дальнейшего анализа.

Например, в предыдущих работах мы неоднократно сталкивались с тем, что ряд алгоритмов (MicrosoftNaiveBayes и др.) требуют предварительной дискретизации непрерывных значений числовых параметров. Но в ряде случаев пользователю желательно посмотреть возможные диапазоны, уточнить их число и т.д. Отдельный интерес может представлять и распределение строк по значению выбранного параметра.

Explore Data

Инструмент Explore Data позволяет проанализировать значения столбца (или диапазона ячеек) и отобразить их на диаграмме. Рассмотрим его работу на примере значения годового дохода клиента (Income). Дополнительный интерес представляет то, что это значение может рассматриваться и как непрерывное, и как дискретное. Итак, запускаем инструмент (рис. 13.2).

1 2



3 4

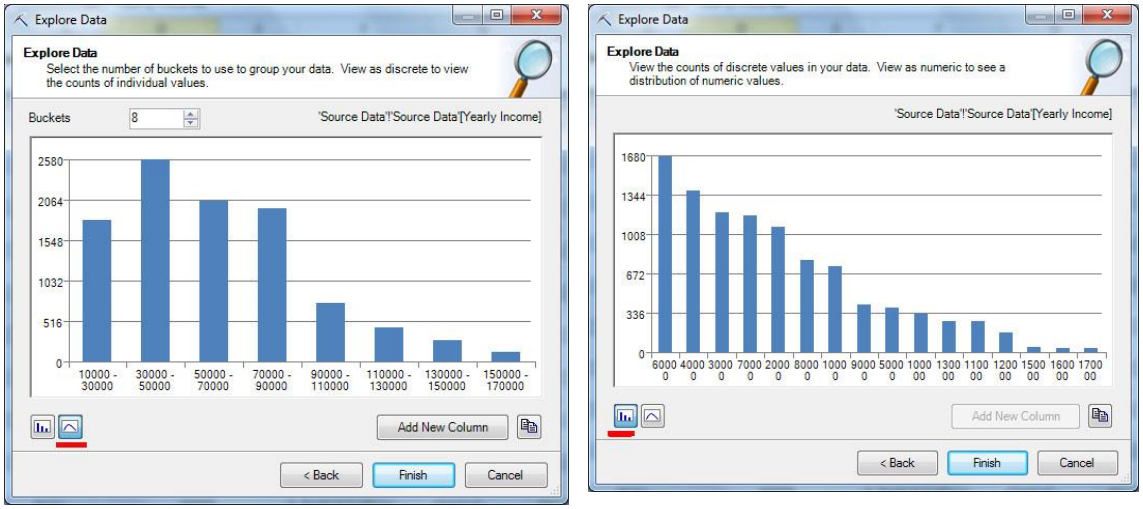


Рис. 13.2. Инструмент Explore Data

В процессе работы потребуется указать, для какой таблицы (или диапазона ячеек) и столбца будет проводиться анализ (рис. 13.2-1 и рис. 13.2-2). После чего указанные значения будут проанализированы и результат представлен в виде гистограммы.

Как уже отмечалось выше, значение годового дохода можно рассматривать и как непрерывное, и как дискретное (за счет того, что в нашем наборе данных присутствуют только значения, кратные 10 тысячам).

Для непрерывного значения будет предложен вариант разбиения на диапазоны (рис. 13.2-3). Число диапазонов можно поменять, и диаграмма с распределением значений будут построена заново. Нажав кнопку "Add New Column" можно добавить в исходную таблицу новый столбец с интервалами годового дохода. Например, если для строки значение Yearly Income = 30000, то значение нового параметра Yearly Income 2 при использовании представленного на рисунке разбиения будет "'30000 - 50000" (именно так, с апострофом в начале, чтобы рассматривалось как строковое). В ходе интеллектуального анализа, полученный столбец может использоваться вместо исходного (включение обоих столбцов одновременно нежелательно).

Кнопками с изображениями графика и гистограммы (на рис. 13.2-3, рис. 13.2-4 они подчеркнуты), можно указать тип анализируемого значения - непрерывное или дискретное. Если значение годового дохода рассматриваем как дискретное, то для него будет построена диаграмма, показывающая распределение числа строк по значению годового дохода (рис. 13.2-4). При этом сортировка производится по убыванию числа строк с данных значением, из-за чего первый столбец гистограммы соответствует значению "60000", второй - "40000" и т.д. Сформированную гистограмму можно скопировать в буфер (кнопка правее кнопки "Add New Column", рис. 13.2-3, рис. 13.2-4) и использовать для дальнейшей работы.

Clean Data

Инструмент Clean Data(рис. 13.3) позволяет подготовить данные для анализа, отбросив нетипичные или ошибочные данные (выбросы), а также проведя замену отдельных значений. Как отмечается в документации, под выбросом подразумевается значение данных, являющееся проблематичным по одной из следующих причин:

- значение находится за пределами ожидаемого диапазона;
- данные были введены неправильно;
- значение отсутствует;
- данные представляют собой пробел или пустую строку;
- значение может значительно отклониться от распределения, которому подчиняются данные в модели.

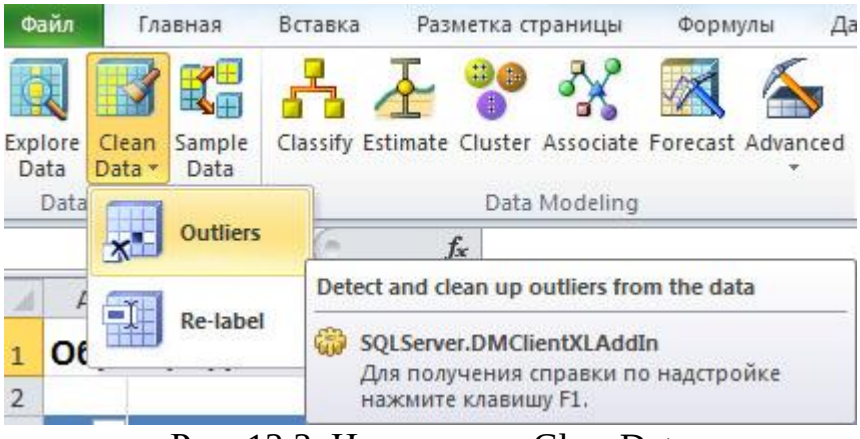


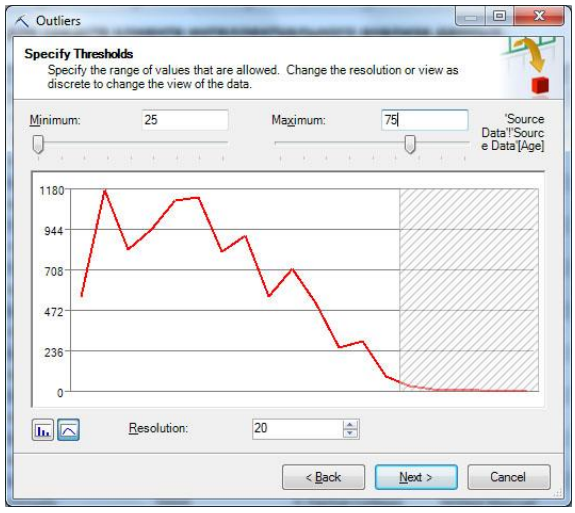
Рис. 13.3. Инструмент CleanData

Использование данного инструмента проиллюстрируем на примере все той же таблицы с данными о клиентах (лист Source Data). Обратимся к столбцу с возрастом. Пусть нам нужно очистить набор данных от информации о нехарактерных по возрасту покупателях. Запускаем инструмент Clean Data->Outliers, в окне аналогичном представленному на рис. 13.2-1 выбираем таблицу для анализа, затем в окне Select Column(рис. 13.2-2)- столбец Age.

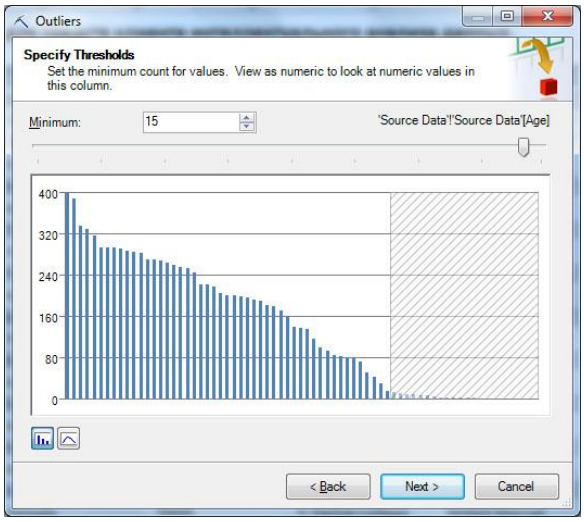
В рассматриваемом наборе данных есть строки со значениями столбца Age от 25 до 96 лет. Если этот параметр считаем непрерывным, то он будет представлен графиком, где по оси X указывается возраст, по оси Y-число клиентов с таким возрастом. В наборе данных доля клиентов преклонного возраста очень мала. На рис. 13.4-1 показано, что установив пороговое значение в 75 лет, мы отбрасываем заштрихованный "хвост", включающий нехарактерные значения (покупатели велосипедов в возрасте от 76 до 96 лет, которых подавляющее меньшинство).

Во многом аналогично выглядит работа с параметром, принимающим дискретные значения. Для него строится гистограмма, а для определения порога нужно указать минимальное число примеров, "поддерживающих" значение. Например, на рис. 13.4-2, установлено пороговое значение в 15. К сожалению, при большом числе столбцов гистограммы, значения параметра на ней не отображаются. Поэтому не понять, что именно попадает в "хвост" распределения.

1



2



3

4

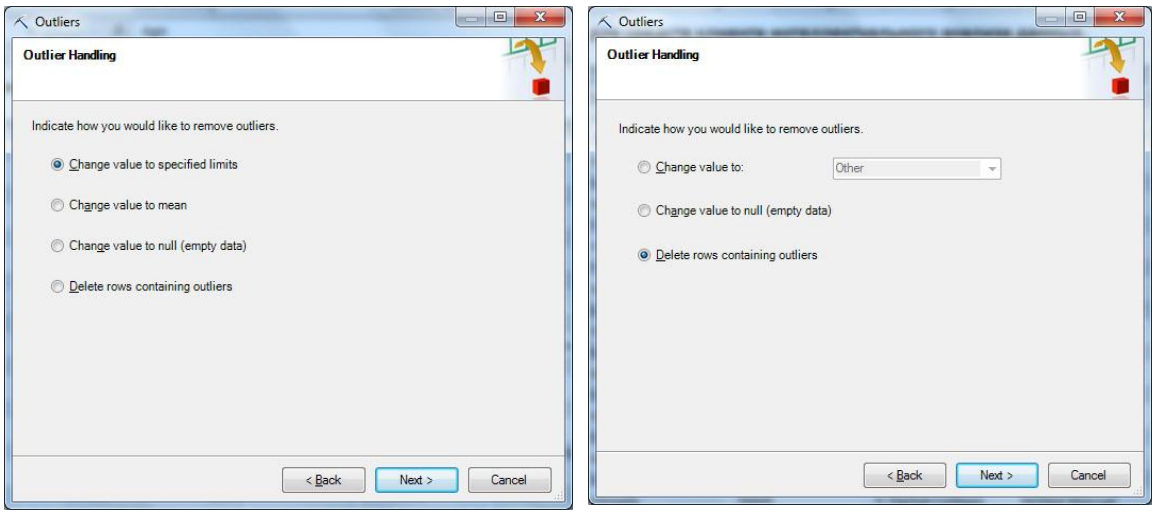


Рис. 13.4. Использование инструмента CleanData для исключения выбросов

Итак, мы выделили нехарактерные данные. Теперь нужно определить, что с ними делать. Предлагаемые мастером решения несколько отличаются для случаев непрерывного и дискретного параметра. Соответствующую строку можно удалить (Delete rows containing outliners) или заменить значение параметра на пустое (Change value to null). Кроме того, для непрерывных данных (рис. 13.2-3) можно заменить нехарактерное значение средним или граничным (сверху или снизу, в зависимости от того, какой диапазон отбрасывается). Для дискретного параметра (рис. 13.2-4) можно указать значение (из числа уже имеющихся в наборе), на которое будут заменяться "выбросы".

Последнее окно мастера (оно на рисунке не представлено) предлагает выбрать, куда заносить изменения - в исходные данные (Change data inplace), в их копию на новом листе Excel (Copy sheet data with changes to a new work sheet) или в новый столбец в исходной таблице (Add as a new column to the current work sheet). Последняя опция для случая удаления строк недоступна.

CleanData.Re-label

В некоторых случаях в исходных данных могут быть значения, которые затрудняют автоматизированный анализ. Например, есть параметр "город" и среди его значений - Санкт-Петербург, С-Петербург, СПб. Для того, чтобы в процессе интеллектуального анализа эти значения учитывались корректно, надо их заменить на одно. Для этого можно использовать инструмент Re-label. Его же можно применить, если требуется снизить уровень детализации значений параметра. Надо отметить, что инструмент работает только с дискретными значениями (ну или рассматриваемыми как дискретные).

Для примера, в таблице с информацией о клиентах нам надо уменьшить число значений параметра CommuteDistance (расстояние ежедневных поездок). Исходные значения "0-1 Miles", "1-2 Miles", "2-5 Miles", "5-10 Miles", "10+ Miles". Пусть все, что меньше 2 миль, будет "близко", остальное

- "далеко". Добавим в таблицу две пустые строки и укажем для одной CommuteDistance "близко», для другой - "далеко". Делается это потому, что значения, на которые заменяем, тоже должны присутствовать в столбце.

1 2

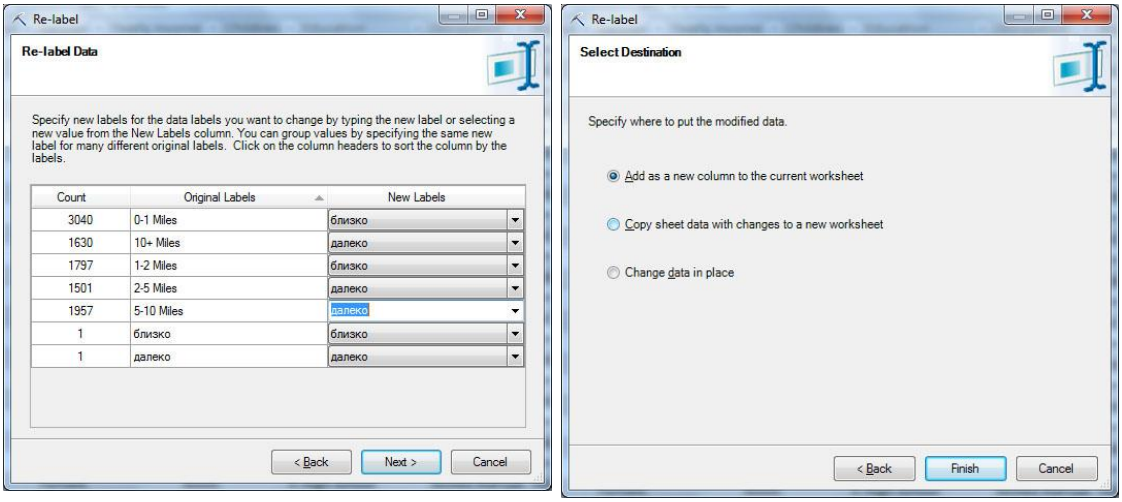


Рис. 13.5. Замена обозначений

Запустим инструмент: CleanData->Re-label. Первые два экрана, как и ранее, позволяют указать таблицу и столбец. Далее указываем порядок замены (рис. 13.5-1) и выбираем создание нового столбца (рис. 13.5-2), чтобы не потерять исходные данные. Замена будет произведена, после чего не забудем удалить добавленные пустые строки с "близко"-"далеко".

SampleData

Последний инструмент в группе Data Preparation называется Sample Data (Образцы данных). Он позволяет решить задачу формирования, обучающего и тестового множеств данных, а также выполнять "балансировку" данных.

В тех случаях, когда используемый метод интеллектуального анализа требует предварительного обучения модели (например, для решения задачи классификации) необходимо сформировать несколько наборов данных - для обучения модели, проверки ее работы, собственно анализа. Инструмент Sample Data позволяет подготовить нужные наборы.

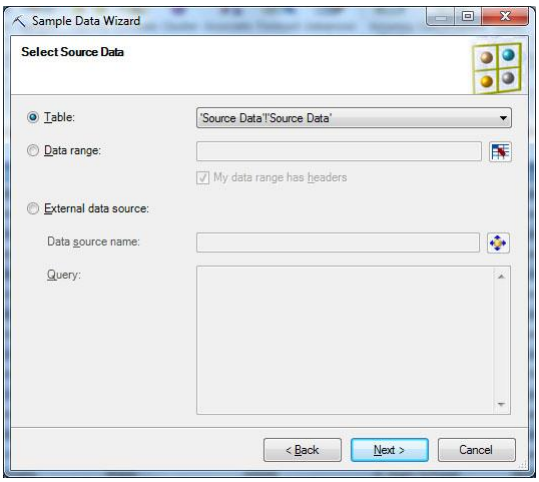
Пусть необходимо случайным образом разделить имеющийся набор данных на обучающую и тестовую выборку. Для этого надо запустить инструмент Sample Data, указать откуда берем данные для обработки (рис. 13.6-1) и тип формируемой выборки. Сначала сделаем случайную выборку, т.е. тип - Random Sampling (рис. 13.6-2). Далее указывается процент записей из исходного набора (или точное число записей) помещаемых в выборку (рис. 13.6-3) и место для сохранения полученных результатов. На рис. 13.6-4 видно, что можно отдельно сохранить сформированную выборку и данные, в нее не попавшие. В итоге можем получить обучающий и тестовый наборы. Хотелось бы обратить внимание на возможность использования внешнего источника данных при формировании выборки (рис. 13.6-1). Это позволяет

использовать данные хранящиеся на MS SQLServer для формирования наборов значений. Но как отмечается в описании инструмента, при использовании внешнего источника данных в окне, представленном на рис. 13.2, будет доступен только параметр случайной выборки.

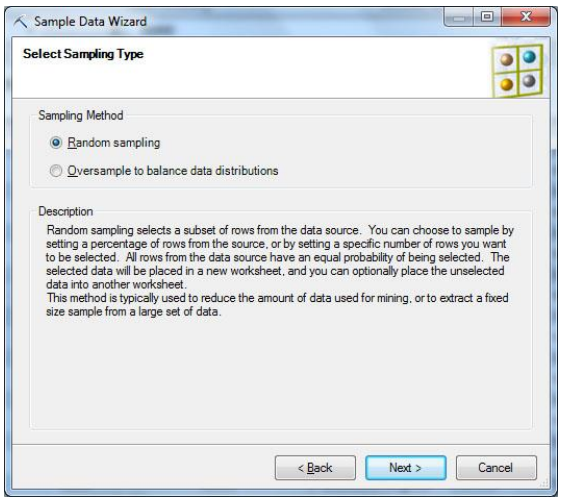
При использовании средств интеллектуального анализа для обнаружения редких событий, в обучающем наборе рекомендуется увеличить частоту появления нужного события по сравнению с исходными данными. Формирование подобной выборки часто называют балансировкой данных, и инструмент SampleData позволяет ее выполнить.

С помощью инструмента Explore Data проанализируем распределение клиентов в наборе данных по регионам. На рис. 13.7-1 видно, что примерно пятая часть клиентов у нас из региона Pacific (будем считать это Азиатско-Тихоокеанским регионом). Сформируем набор данных, где таких клиентов будет 50 %.

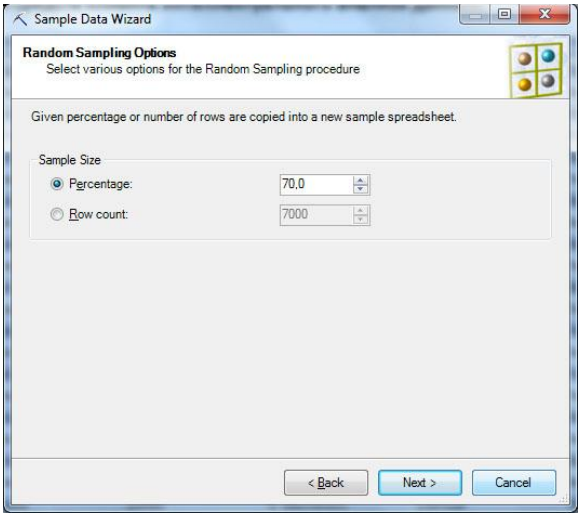
1



2



3



4

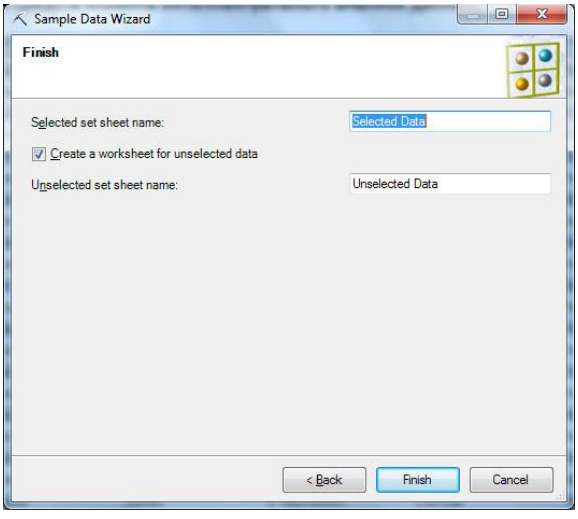
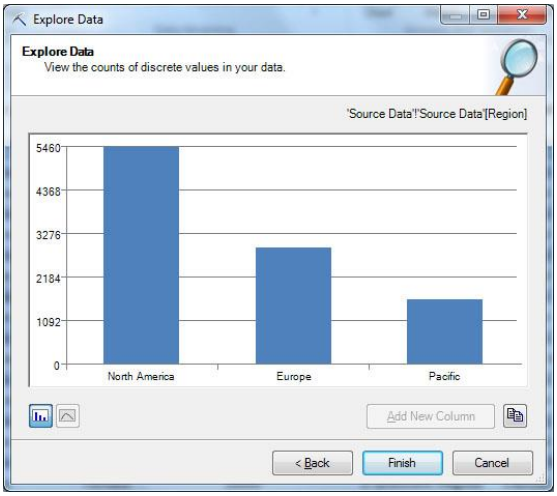
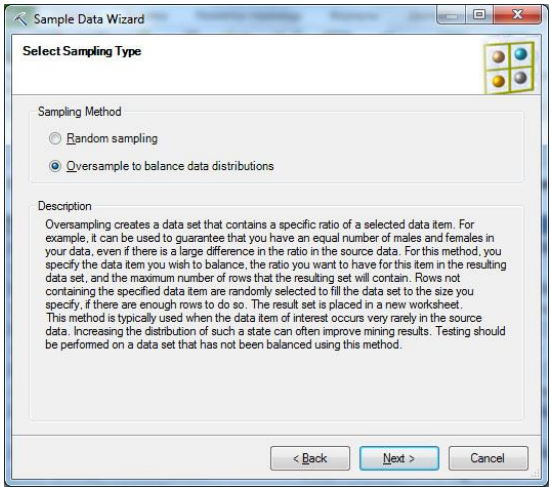


Рис. 13.6. Инструмент Sample Data

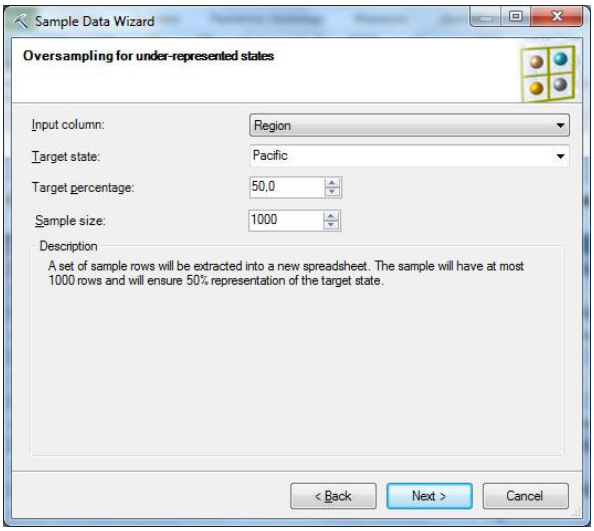
Запустим инструмент Sample Data, укажем в качестве источника данных используемую таблицу Excel и выберем вариант формирования избыточной выборки с балансировкой данных (Oversample to balance data distributions, рис. 13.7-2). Далее укажем столбец, для которого выполняется балансировка, и частоту появления нужного значения и размер выборки (рис. 13.7-3).Будет создана новая таблица с указанным пользователем названием. Снова применим Explore Data и убедимся в том, что выборка сформирована в соответствии с указанными выше требованиями (рис. 13.7-4).

1 2





3 4



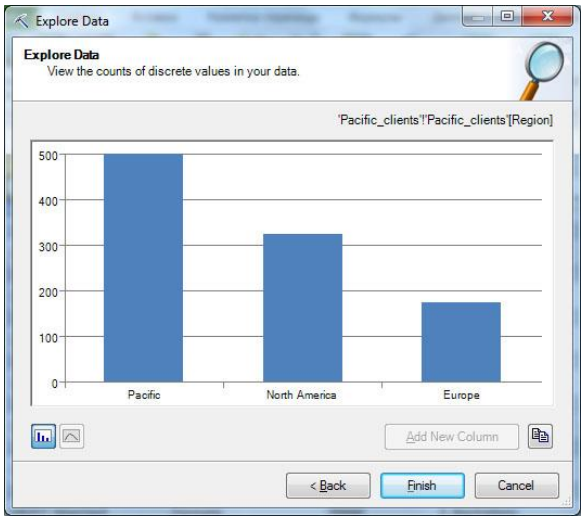


Рис. 13.7. Формирование выборки с заданным распределением клиентов по регионам

Задание. Проведите описанную в лабораторнойобработку выбранного набора данных.

Лабораторная работа 7. Использование инструментов Data Mining Client для Excel для создания модели интеллектуального анализа данных.

Цель: в лабораторной работе будет рассмотрен процесс создания модели интеллектуального анализа с помощью инструментов, входящих в состав Data Mining Client для Excel.

Рассмотренные в лабораторных работах "Надстройки интеллектуального анализа данных для MicrosoftOffice" - "Использование инструментов "Prediction Calculator" и "ShoppingbasketAnalysis"" "Средства анализа таблиц для Excel" (TableAnalysisTools) для конечного пользователя во многом представляются "черным ящиком", выполняющим анализ,но не дающим информации о том, как получен результат. Если такое решение не устраивает, можно перейти с вкладки Analyze на вкладку DataMining и воспользоваться инструментами DataMiningClient для Excel (рис. 14.1).

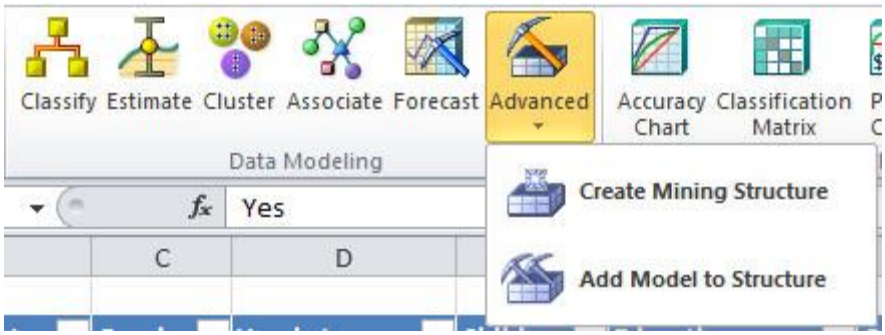


Рис. 14.1. Группа инструментов DataModeling

В "Использование инструментов Data Mining Client для Excel 2007 для подготовки данных" мы рассмотрели инструменты, позволяющие подготовить данные для анализа. Следующая группа показанные на рис. 14.1 инструменты DataModeling,позволяющие создать модели интеллектуального анализа данных.

Классификация (Classify)	**создает модель классификации на основе существующих данных таблицы Excel, диапазона Excel или внешнего источника данных (AnalysisServicesDataSource). На основе обрабатываемых данных формируются шаблоны, которые при использовании позволяют отнести рассматриваемый пример к одному из возможных классов. По умолчанию используется алгоритм DecisionTrees, но также доступны LogisticRegression, NaiveBayes, NeuralNetworks.**
Оценка (Estimate)	**позволяет создать модель оценки значения целевого параметра (он должен быть числовым) на основе данных из таблицы или диапазона ячеек Excel либо**

	внешнего источника данных. По умолчанию используется алгоритм Decision Trees, также доступны Linear Regression, Logistic Regression, Neural Networks.
Кластер (Cluster)	**запускает мастер, позволяющий построить модель кластеризации на основе данных из таблицы или диапазона Excel, либо внешнего источника данных. Модель определяет группы строк со сходными характеристиками,для чего используется алгоритм MicrosoftClustering. Данная задача аналогична решаемой средством DetectCategories из набора TableAnalysisTools.**
Поиск взаимосвязей (Associate)	**помогает создать модель, описывающую взаимосвязь объектов (покупаемых товаров и т.д.), затрагиваемых одной транзакцией, для чего используется алгоритм AssociationRules. С подобной задачей мы сталкивались, используя инструмент ShoppingBasketAnalysis из TableAnalysisTools. Для построения модели анализа необходимо, чтобы исходные данные содержали столбец с идентификатором транзакций и были по нему отсортированы. В качестве источника данных может использоваться только таблица или диапазон ячеек Excel.**
Прогноз (Forecast)	**Данный мастер позволяет построить модель для прогнозирования новых значений в числовой последовательности, аналогично инструменту Forecast в TableAnalysisTools. Используется алгоритм TimeSeries, для работы которого требуется, чтобы столбец (или столбцы), в отношении которого будет выполняться прогноз, имели непрерывные числовые значения. Также может присутствовать столбец с отметкой времени (в этом случае, строки в таблице должны быть по нему отсортированы).**
Дополнительно (Advanced)	**позволяет создать структуру1 интеллектуального анализа данных или добавить в существующую структуру новую модель (например, для сравнения результатов, выдаваемых разными алгоритмами анализа).**

Используем инструмент Classify. В поставляющемся с надстройками наборе данных (меню "Пуск"->"Надстройки интеллектуального анализа данных"->"Образцы данных Excel") выберем таблицу TrainingData,содержащую случайную выборку 70% данных из таблицы SourceData. Запустим мастер Classify, в первом окне которого будет

комментарий по применению инструмента, а второе окно позволит указать источник данных для анализа (таблица TrainingData). Дальше потребуется описать цель анализа.

Пусть нас интересует, сделает ли данный клиент покупку. В целевом столбце указываем параметр BikeBuyer (рис. 14.2, окно слева), сбрасываем в перечне входных столбцов отметки напротив ID (порядковый номер клиента в базе никак не влияет на его решение о покупке). Если ID оставить среди анализируемых параметров, то итоговая модель может его учесть. В частности, на рис. 14.3 показано дерево решений, учитывающее значение поля ID в процессе классификации, что однозначно неправильно.

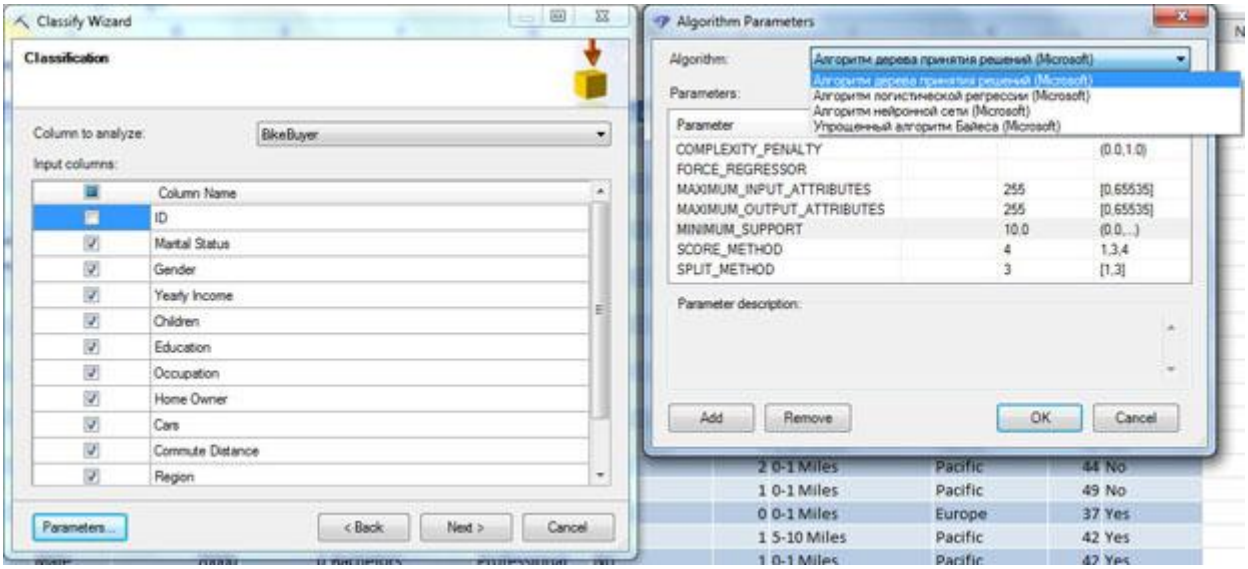


Рис. 14.2. Указание параметров для анализа

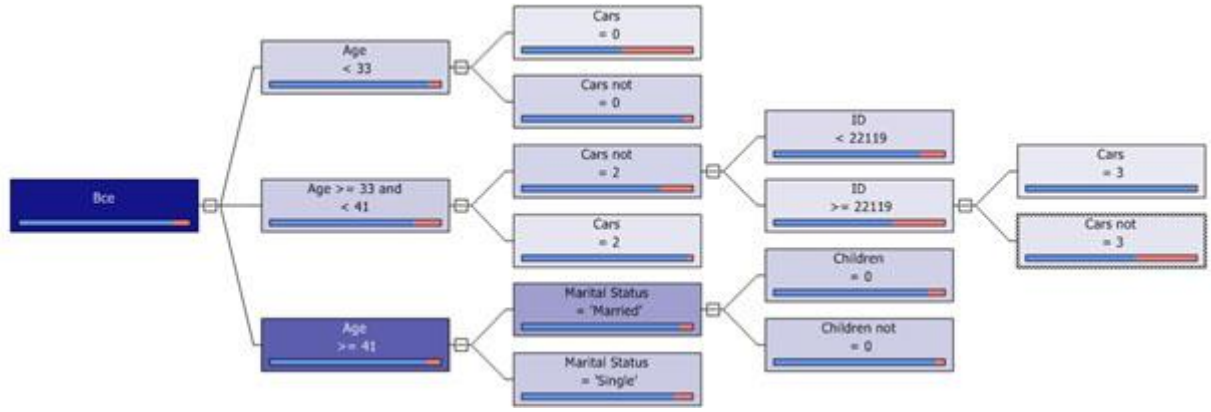


Рис. 14.3. Неудачный вариант дерева решений

Если требуется более точная настройка, можно открыть окно Parameters и явно указать используемый алгоритм и его параметры (рис. 14.2, окно справа). Далее мастер предложит разделить имеющиеся данные на набор для обучения модели и для ее тестирования (рис. 14.4-1). По умолчанию на набор для тестирования выделяется 30 % строк исходного набора.

1 2

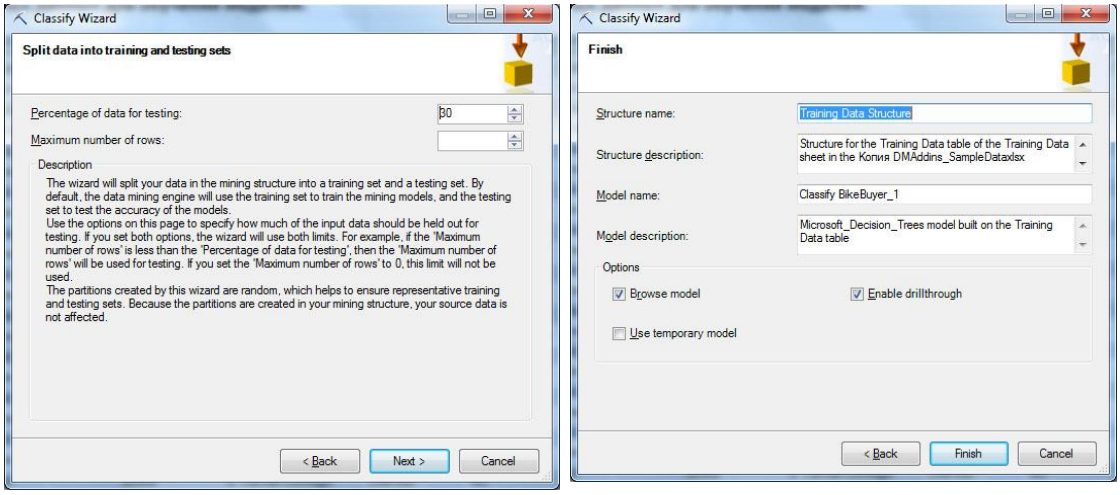


Рис. 14.4. Разбиение данных и указание названий модели и структуры

Последний этап работы мастера - указание имени создаваемой структуры и модели (рис. 14.4-2). В нашем примере структура будет называться TrainingDataStructure, а модель Classify BikeBuyer_1. Эти названия нам понадобятся впоследствии для работы с моделью.

Если выполняющий анализ пользователь не имеет прав администратора в базе Аналитических Служб (эту настройку мы делали в "Надстройки интеллектуального анализа данных для MicrosoftOffice"), то создать постоянною модель интеллектуального анализа на сервере он не сможет. В этом случае можно использовать временную модель, для чего отметить пункт Use temporary model. Временная модель будет автоматически удалена с сервера поз авершению сеанса работы пользователя.

Отмеченная по умолчанию настройка Brose model указывает на то, что после создания модели будет открыто окно просмотра. Для модели, созданной с использованием алгоритма DecisionTrees, отображается построенное дерево решений и диаграмма зависимостей. Представленное на рис. 14.5 дерево решений позволяет оценить построенную модель. Расположенные в верней части экрана "ползунок" и выпадающий список позволяют установить число отображаемых уровней дерева (на рисунке показаны все пять). Если навести указатель мыши на точку ветвления, можно увидеть всплывающую подсказку с указанием того, сколько и каких случаев в обучающем наборе ей соответствует. Для выделенного узла в правой части экрана отображается его описание и гистограмма с распределением значений. Кнопкой Copy to Excel можно перенести результат из окна просмотра на новый лист Excel (для дерева решений в Excel будет перенесено его растровое изображение).

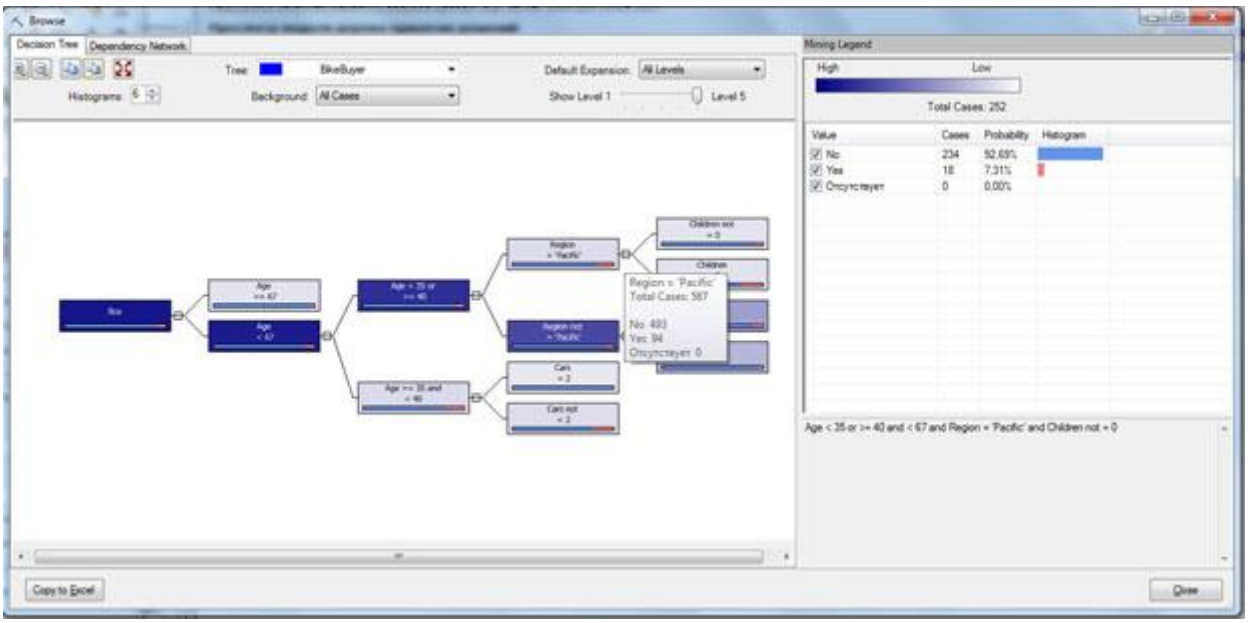
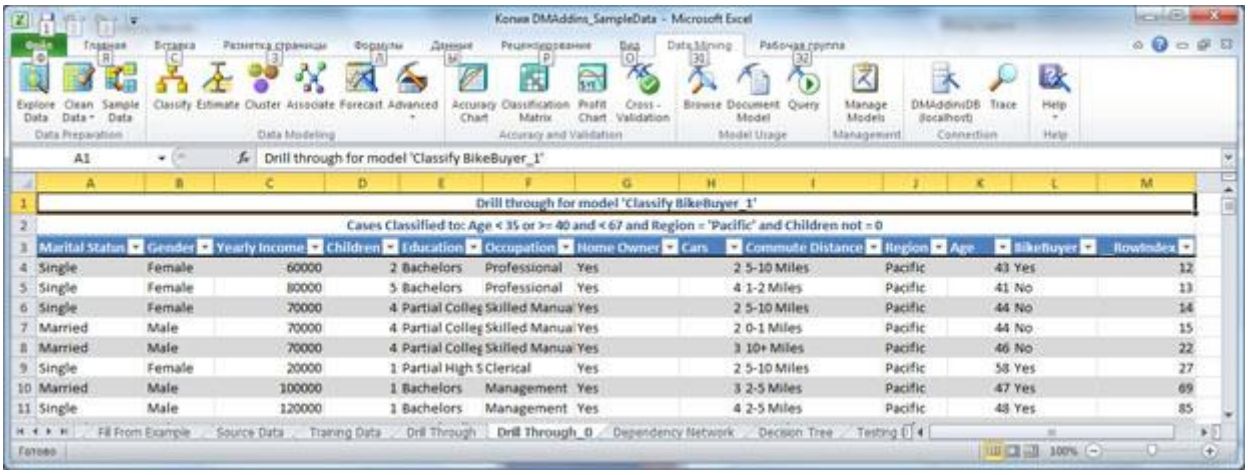


Рис. 14.5. Построенное дерево решений

Щелкнув по узлу дерева правой клавишей мыши и выбрав в контекстном меню DrillThroughModelColumns (можно примерно перевести как "детализация использовавшихся моделью данных") мы получим новую таблицу Excel, содержащую набор строк из обучающей выборки, которые соответствуют данному узлу (рис. 14.6).



Drill through for model 'Classify BikeBuyer_1'

Cases Classified to: Age < 35 or >= 40 and < 67 and Region = 'Pacific' and Children not = 0

Marital Status	Gender	Yearly Income	Children	Education	Occupation	Home Owner	Cars	Commute Distance	Region	Age	BikeBuyer	RowIndex
Single	Female	60000	2	Bachelors	Professional	Yes	2	5-10 Miles	Pacific	43	Yes	12
Single	Female	80000	5	Bachelors	Professional	Yes	4	1-2 Miles	Pacific	41	No	13
Single	Female	70000	4	Partial Colleg	Skilled Manual	Yes	2	5-10 Miles	Pacific	44	No	14
Married	Male	70000	4	Partial Colleg	Skilled Manual	Yes	2	0-1 Miles	Pacific	44	No	15
Married	Male	70000	4	Partial Colleg	Skilled Manual	Yes	3	10+ Miles	Pacific	46	No	22
Single	Female	20000	1	Partial High S	Clerical	Yes	2	5-10 Miles	Pacific	58	Yes	27
Married	Male	100000	1	Bachelors	Management	Yes	3	2-5 Miles	Pacific	47	Yes	69
Single	Male	120000	1	Bachelors	Management	Yes	4	2-5 Miles	Pacific	48	Yes	85

Рис. 14.6. Результат выполнения DrillThroughModelColumns

На рис. 14.7 представлена диаграмма зависимостей, показывающая выявленные взаимосвязи между параметрами. Ее также можно скопировать в Excel. Выделяя на диаграмме узел, можно увидеть все влияющие на него.

Закроем окно просмотра модели. Если нужно будет снова просмотреть ее параметры, воспользуйтесь инструментом Browse, который находится в группе ModelUsage.

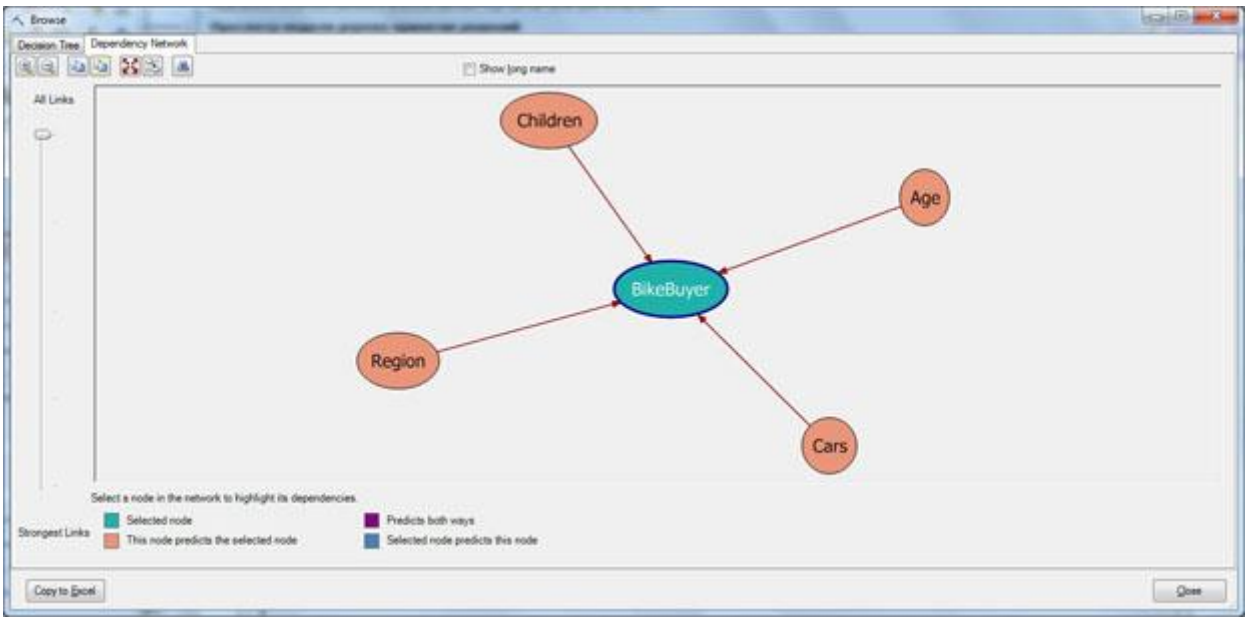


Рис. 14.7. Диаграмма зависимостей

Для того чтобы управлять имеющимися на сервере структурами и моделями интеллектуального анализа, можно воспользоваться соответствующим мастером, запускаемым по нажатию кнопки ManageModels на вкладке DataMining (рис. 14.8). Он позволяет просмотреть имеющиеся структуры и модели, переименовать их, удалить ненужные, выполнить другие действия на сервере прямо из DataMiningClient.

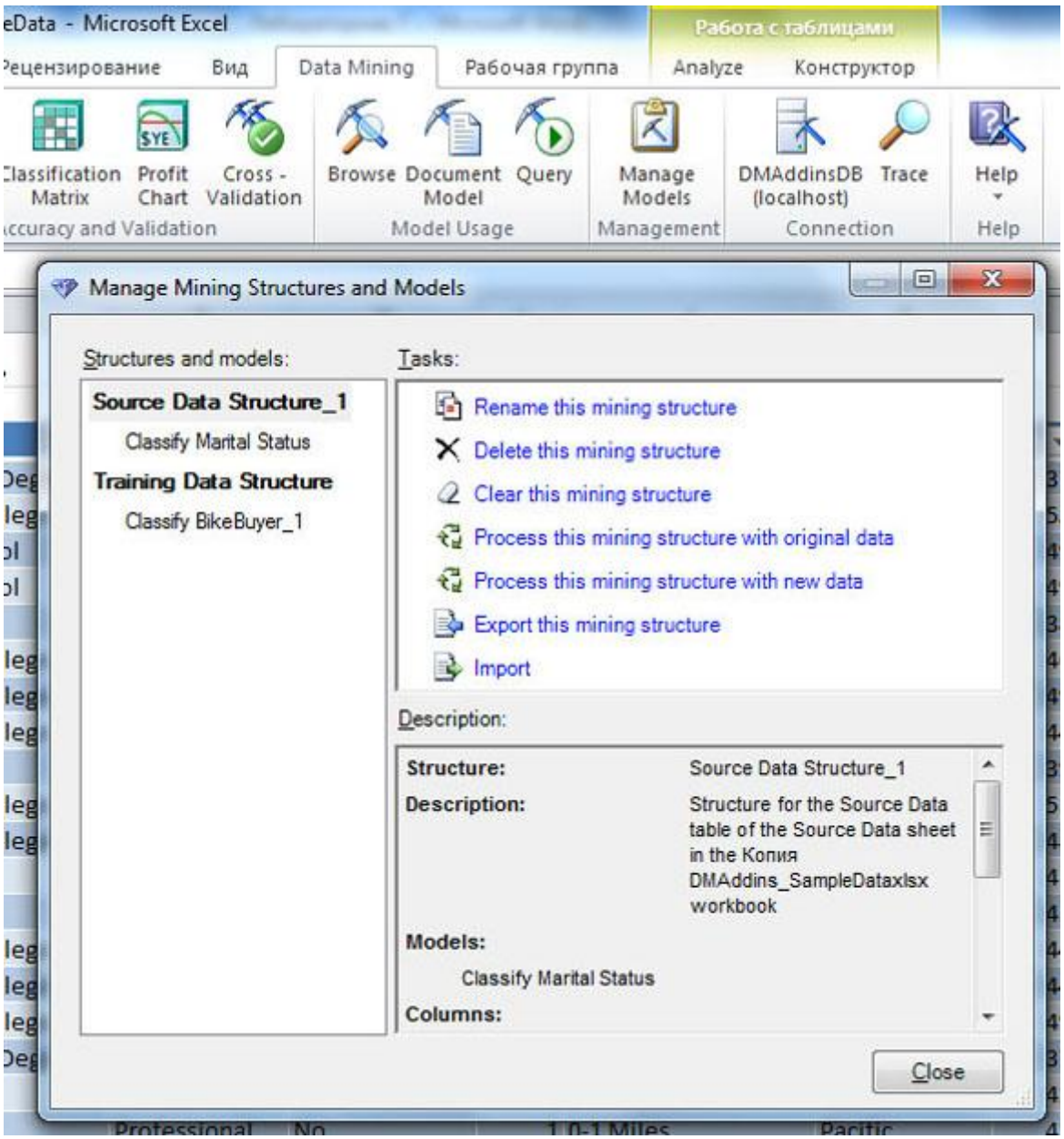


Рис. 14.8. Окно управления моделями

Задание 1. Создайте модель интеллектуального анализа, аналогичную описанной в лабораторной работе.

Задание 2. Воспользуйтесь набором данных в таблице Associate и одноименным мастером для создания модели, описывающей взаимосвязи между категориями товаров в одном заказе. При необходимости, воспользуйтесь справочной системой по инструменту. Проанализируйте выявленные правила и диаграмму зависимостей. Сравните с результатами, полученными в "Использование инструментов "Prediction Calculator" и "ShoppingbasketAnalysis""(раздел "Анализ покупательского поведения").

Лабораторная работа 8. Анализ точности прогноза и использование модели интеллектуального анализ

Цель: Лабораторная работа посвящена проверке точности модели и выполнению запросов к модели интеллектуального анализа.

В предыдущей лабораторной работе мы создали модель для классификации клиентов магазина с целью определить, сделает ли данный клиент покупку или нет. Следующая задача - оценить точность построенной модели интеллектуального анализа. Для этого можно использовать инструменты из группы Accuracy and Validation (в русском варианте - Точность и Правильность).

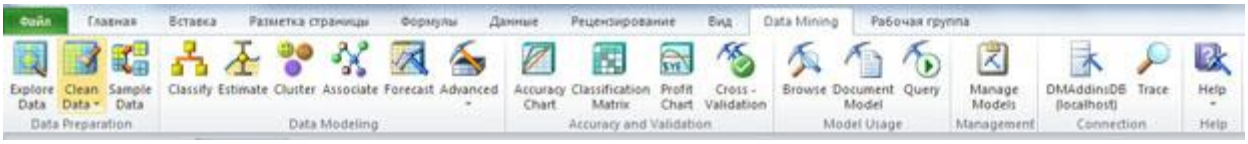


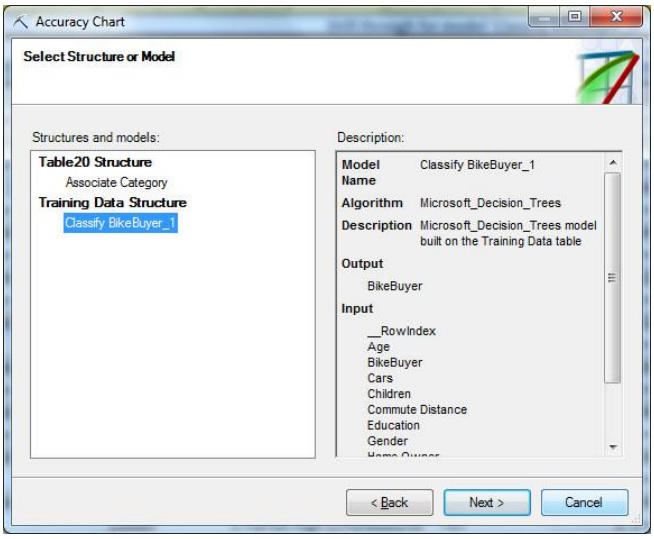
Рис. 15.1. Инструменты DataMiningClient

Диаграмма точности (AccuracyChart) позволяет, применив модель на тестовой выборке данных, оценить результаты ее работы. В ходе выполнения "Использование инструментов Data Mining Client для Excel 2007 для создания модели интеллектуального анализа данных" была создана структура TrainingDataStructure и модель классификации Classify BikeBuyer_1. При создании модели мы резервировали 30% данных для целей тестирования (рис. 15.4-1 в "Использование инструментов Data Mining Client для Excel 2007 для создания модели интеллектуального анализа данных").

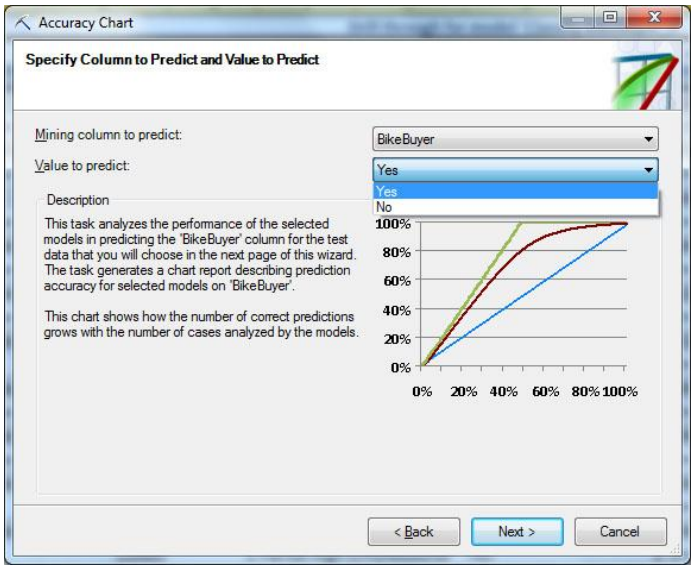
Запустим инструмент AccuracyChart.Первое окно мастера содержит краткое описание инструмента, в следующем - надо указать структуру или модель (рис. 15.2-1). Если для одной структуры определены несколько моделей, по диаграмме можно будет провести их сравнительный анализ. Следующее окно (рис. 15.2-2) служит для выбора предсказываемого параметра и его значения. В нашем случае параметр - BikeBuyer, а оценивать будем точность предсказания значения "Yes". Дальше требуется указать источник данных для тестирования. Это могут быть зарезервированные при создании модели данные, данные из таблицы или диапазона ячеек Excel, или из внешнего источника (рис. 15.2-3). Сейчас выберем данные из модели. В случае указания таблицы Excel (что будем делать в упражнениях), надо описать соответствие столбцов в модели и используемой для тестирования таблице (рис. 15.2-4). После этого будут сформированы и помещены на новый лист Excel диаграмма точности (рис. 15.3) и таблица со значениями, представленными на диаграмме (рис. 15.4).

На диаграмме красная линия соответствует идеальной модели, светло-зеленая - нашей модели, нижняя (синяя) линия - линия случайного выбора, всегда идет под углом 45 градусов.

1.


Accuracy Chart
Select Structure or Model
Structures and models:
Table20 Structure
Associate Category
Training Data Structure
Classify BikeBuyer_1
Description:
Model Name
Classify BikeBuyer_1
Algorithm
Microsoft_Decision_Trees
Description
Microsoft_Decision_Trees model built on the Training Data table
Output
BikeBuyer
Input
__RowIndex
Age
BikeBuyer
Cars
Children
Commute Distance
Education
Gender
< Back
Next >
Cancel

2.


Accuracy Chart
Specify Column to Predict and Value to Predict
Mining column to predict:
BikeBuyer
Value to predict:
Yes
Yes
No
Description
This task analyzes the performance of the selected models in predicting the 'BikeBuyer' column for the test data that you will choose in the next page of this wizard. The task generates a chart report describing prediction accuracy for selected models on 'BikeBuyer'.
This chart shows how the number of correct predictions grows with the number of cases analyzed by the models.
100%
80%
60%
40%
20%
0%
0% 20% 40% 60% 80% 100%
< Back
Next >
Cancel

3.

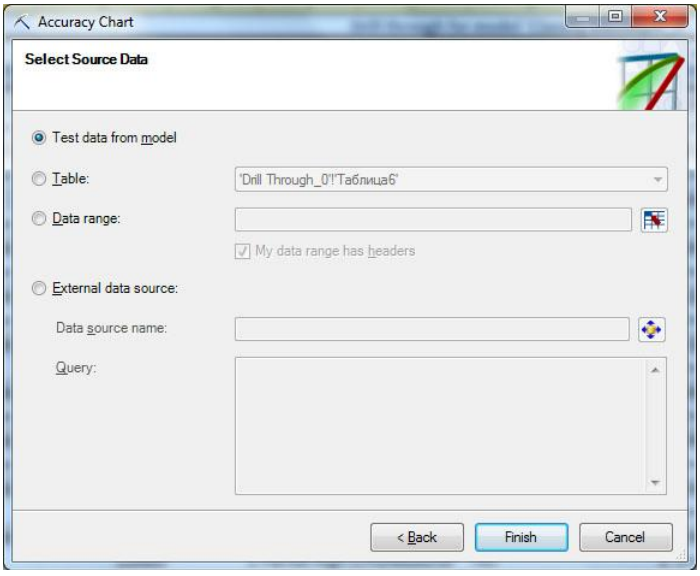
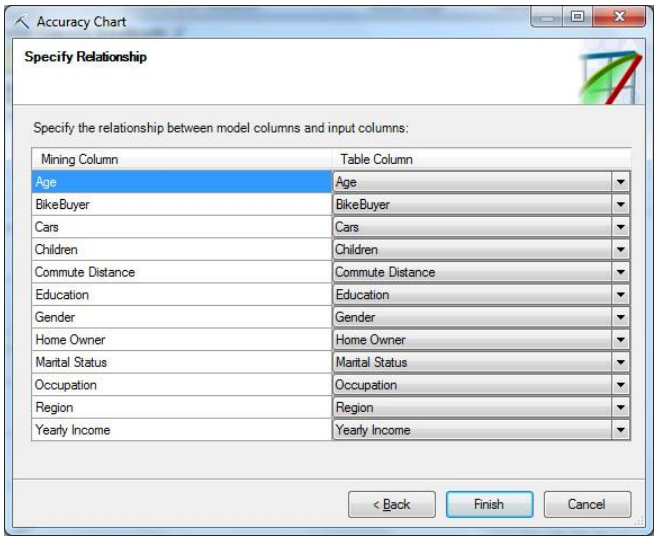

Accuracy Chart
Select Source Data
Test data from model
Table:
'Drill Through_0'!'Таблица6'
Data range:
My data range has headers
External data source:
Data source name:
Query:
< Back
Finish
Cancel

Рис. 15.2. Мастер построения диаграммы точности

4.



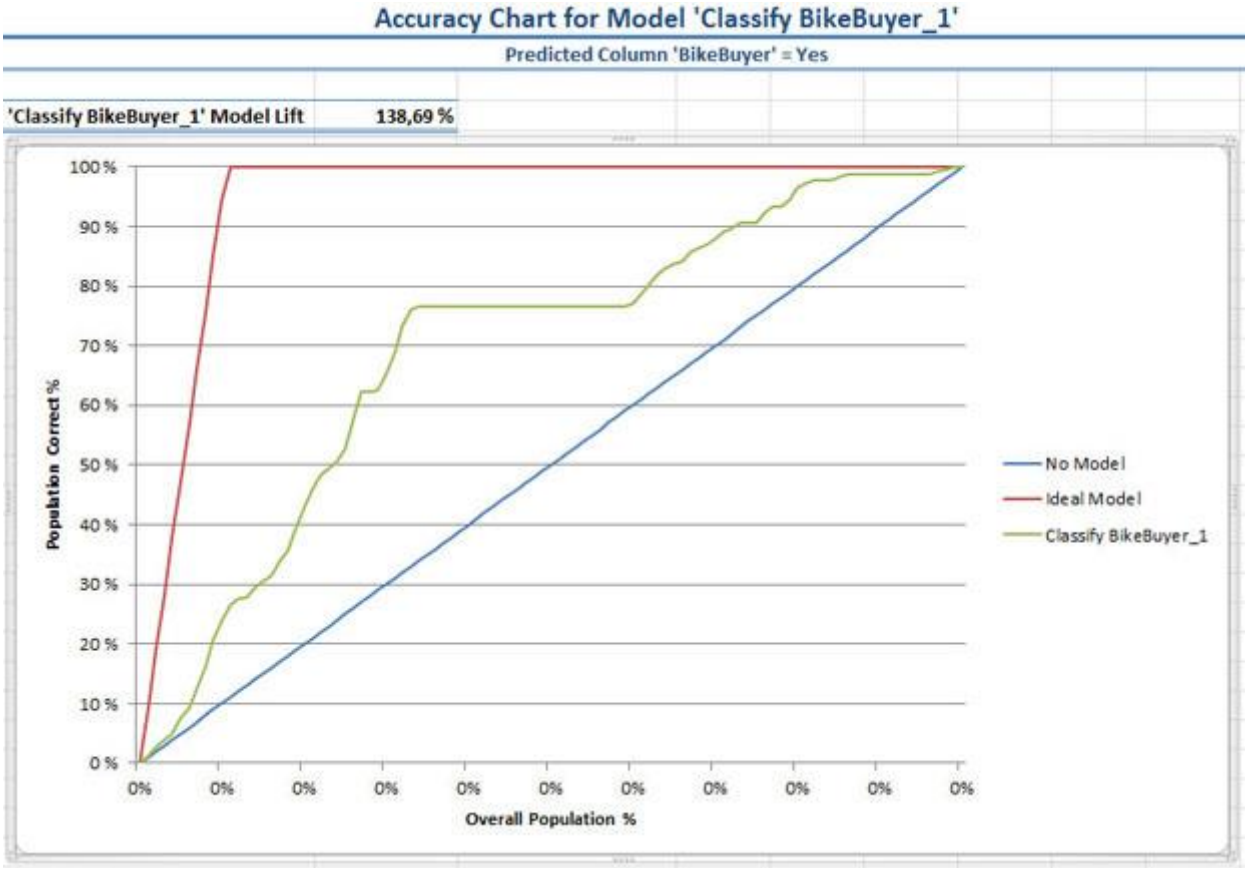


Рис. 15.3. Диаграмма точности (AccuracyChart)

Данные на диаграмме и в таблице можно интерпретировать следующим образом. Пусть нам необходимо выбрать всех клиентов, которые сделают покупки. Формируемая идеальной моделью выборка объемом в 11% от числа исходных записей будет включать все 100% нужных записей (в тестовом множестве их видимо чуть меньше 11%). Случайная выборка объемом в 11%

содержит 11% нужных записей, а выборка такого же объема, формируемая нашей моделью - 26,58%. В выборку в 25% от общего объема данных, наша модель поместит 52,7% "правильных" клиентов и т.д. Качество прогноза падает (горизонтальный участок зеленого графика) после обнаружения 76% интересующих случаев. При визуальном анализе - чем ближе график оцениваемой модели к идеальному, тем более точный прогноз она выдает.

Percentile	Ideal Model	Classify BikeBuyer_1
0 %	0,00 %	0,00 %
1 %	9,46 %	1,35 %
2 %	18,92 %	2,70 %
3 %	28,38 %	4,05 %
4 %	37,84 %	4,95 %
5 %	47,30 %	7,66 %
6 %	56,76 %	9,46 %
7 %	66,22 %	12,61 %
8 %	75,68 %	16,22 %
9 %	85,14 %	20,72 %
10 %	94,59 %	23,87 %
11 %	100,00 %	26,58 %
12 %	100,00 %	27,48 %
13 %	100,00 %	27,93 %
14 %	100,00 %	29,28 %
15 %	100,00 %	30,63 %
16 %	100,00 %	31,53 %
17 %	100,00 %	33,78 %
18 %	100,00 %	35,59 %
19 %	100,00 %	39,19 %
20 %	100,00 %	42,79 %
21 %	100,00 %	45,95 %
22 %	100,00 %	48,20 %
23 %	100,00 %	49,55 %
24 %	100,00 %	50,45 %
25 %	100,00 %	52,70 %

Рис. 15.4. Фрагмент таблицы с оценками точности прогноза

Задание 1. Постройте диаграмму точности аналогичную той, что представлена выше (используемый файл - "Образцы данных Excel"). Дополнительно постройте диаграмму для BikeBuyer = "No". Объясните различие во внешнем виде графиков.

Задание 2. В предыдущем задании для целей тестирования использовались данные из модели. Модель формировалась на данных из таблицы TrainingData. В таблице TestingData находятся 30% данных из исходного набора SourceData. Проверьте точность модели на наборе TestingData.

Анализируя график на рис. 15.3 можно предположить, что у нас с моделью все хорошо. Но обратимся к еще одному инструменту анализа точности - ClassificationMatrix (Матрица классификации). С его помощью мы получаем таблицу с результатами точных и ошибочных предсказаний (рис. 15.5). Из нее видно, что созданная нами модель при тестировании на зарезервированных данных сделала 89,43% правильных прогнозов, что можно расценить как успех (потому и диаграмма точности на рис. 15.3 выглядит хорошо).Но при этом в 100% случаев правильно предсказывала значение "No" и ошибочно "Yes". Иначе говоря, во всех случаях 100% ставится "No". И использовать такую модель для предсказания бессмысленно.

Counts of correct/incorrect classification for model 'Classify BikeBuyer_1'

Predicted Column 'BikeBuyer'

Columns correspond to actual values

Rows correspond to predicted values

Model name:	Classify BikeBuyer_1	Classify BikeBuyer_1
Total correct:	89,43 %	1878
Total misclassified:	10,57 %	222

Results as Percentages for Model 'Classify BikeBuyer_1'

	No(Actual)	Yes(Actual)
No	100,00 %	100,00 %
Yes	0,00 %	0,00 %
Correct	100,00 %	0,00 %
Misclassified	0,00 %	100,00 %

Results as Counts for Model 'Classify BikeBuyer_1'

	No(Actual)	Yes(Actual)
No	1878	222
Yes	0	0
Correct	1878	0
Misclassified	0	222

Рис. 15.5. Матрица классификации

Задание 3. Постройте матрицу классификации, проанализируйте полученный результат.

Проблема, с которой мы столкнулись, могла быть выявлена и раньше, если внимательно посмотреть на построенное дерево решений (рис. 15.6). Но тогда не удалось бы продемонстрировать возможности DataMining по оценке точности модели.

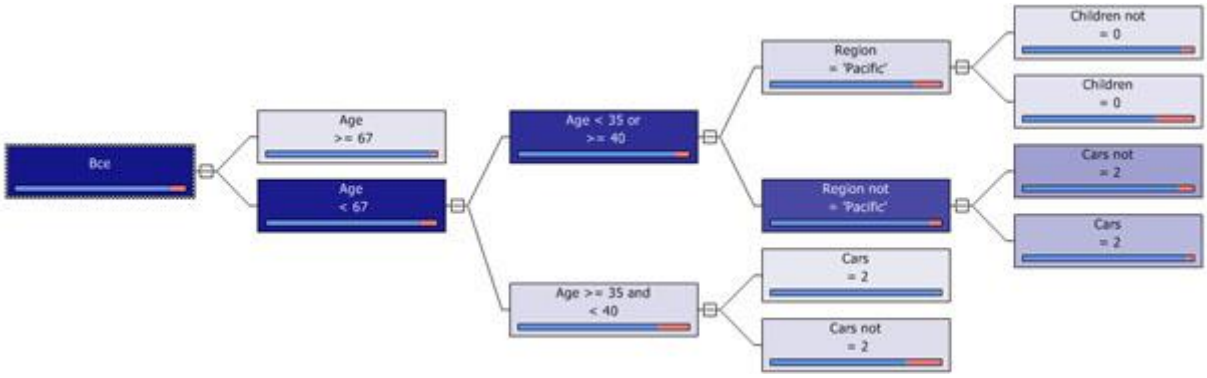


Рис. 15.6. Дерево решений

Из рис. 15.6 видно, что все конечные узлы дерева дают решение BikeBuyer = "No" (ему соответствует синяя полоска на диаграмме характеризующей распределение ответов в обучающей выборке). Ответу "Yes" соответствует более короткая красная полоска, что говорит о том, что поддерживающих такой результат примеров было меньше. По всей видимости, это связано с тем, что таких примеров и вообще меньшинство в рассматриваемом наборе (около 10%).

Попробуем использовать обучающий набор большего объема и с более часто встречающимся значением BikeBuyer = "Yes". Откроем таблицу SourceData, где данных больше. Но процент интересующих нас записей остается таким же (это можно определить с помощью инструмента ExploreData).Поэтому воспользуемся инструментом SampleData ("Использование инструментов Data Mining Client для Excel 2007 для подготовки данных"), чтобы сформировать "избыточную" выборку из 2000 строк, где в 30% случаев BikeBuyer = "Yes". У полученного набора оставим автоматически назначенное название SampledData. С помощью инструмента Classify построим модель аналогично тому, как это было сделано в "Использование инструментов Data Mining Client для Excel 2007 для создания модели интеллектуального анализа данных" (алгоритм - DecisionTrees, целевой параметр BikeBuyer, столбец ID при анализе не учитываем, остальные настройки по умолчанию). Полученное дерево решений представлено на рис. 15.7. Оно проще предыдущего, но в зависимости от значений параметров может давать как прогноз "Yes", так и "No". "Yes" будет в том случае, если у клиента 0 машин и он из региона "Pacific".

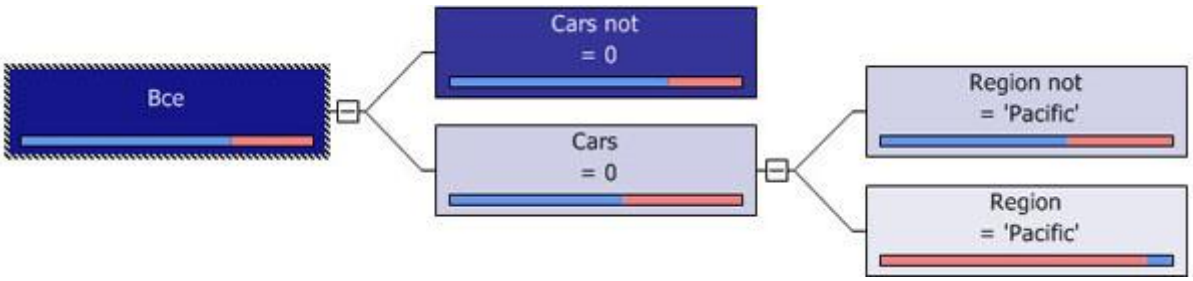
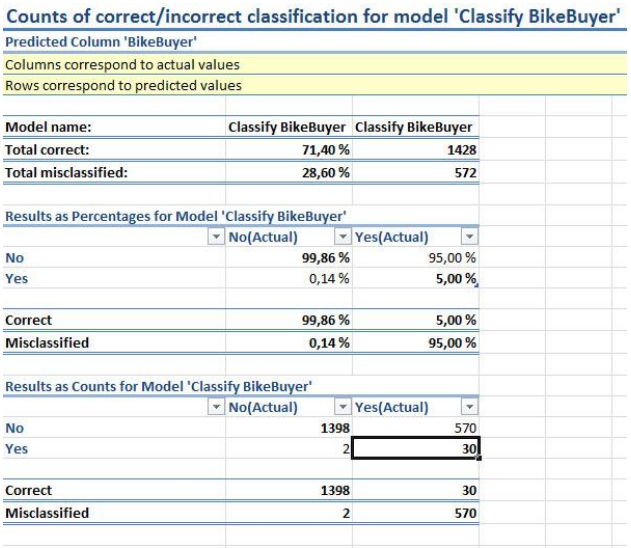


Рис. 15.7. Новое дерево решений

На основе нового набора данных также создадим модель для классификации, основанную на алгоритме NeuralNetworks (нейронных сетей). Если построить для них матрицы классификации (рис. 15.8-1, рис. 15.8-2) будет видно, что модель на основе нейронных сетей дает более точный прогноз. Рассмотренный пример показывает, что в некоторых случаях точность прогноза можно повысить за счет специальной подготовки обучающей выборки и выбора наиболее подходящего алгоритма. Хотя, учитывая относительно высокий процент ошибок, ни та, ни другая модель, наверное, не может быть признана удачной.

1.



Counts of correct/incorrect classification for model 'Classify BikeBuyer'

Predicted Column 'BikeBuyer'

Columns correspond to actual values

Rows correspond to predicted values

Model name:	Classify BikeBuyer	Classify BikeBuyer
Total correct:	71,40 %	1428
Total misclassified:	28,60 %	572

Results as Percentages for Model 'Classify BikeBuyer'

	No(Actual)	Yes(Actual)
No	99,86 %	95,00 %
Yes	0,14 %	5,00 %
Correct	99,86 %	5,00 %
Misclassified	0,14 %	95,00 %

Results as Counts for Model 'Classify BikeBuyer'

	No(Actual)	Yes(Actual)
No	1398	570
Yes	2	30
Correct	1398	30
Misclassified	2	570

Рис. 15.8. Матрицы классификации для дерева решений (1) и нейронных сетей (2)

2.

Counts of correct/incorrect classification for model 'Classify BikeBuyer_3'

Predicted Column 'BikeBuyer'

Columns correspond to actual values

Rows correspond to predicted values

Model name:	Classify BikeBuyer_3	Classify BikeBuyer_3
Total correct:	75,67 %	7567
Total misclassified:	24,33 %	2433

Results as Percentages for Model 'Classify BikeBuyer_3'

	No(Actual)	Yes(Actual)
No	78,70 %	51,60 %
Yes	21,30 %	48,40 %
Correct	78,70 %	48,40 %
Misclassified	21,30 %	51,60 %

Results as Counts for Model 'Classify BikeBuyer_3'

	No(Actual)	Yes(Actual)
No	7083	516
Yes	1917	484
Correct	7083	484
Misclassified	1917	516

Задание. Проведите описанные в работе действия. Прокомментируйте результаты.

Запросы к модели DM

Теперь перейдем к самому интересному - построению запроса к модели интеллектуального анализа. Итак, на сервере есть модель, признанная пригодной для прогнозирования. В используемом нами файле Excel с данными для интеллектуального анализа есть таблица NewCustomers с информацией о новых клиентах (рис. 15.9). В ней есть все столбцы, которые были в наборе SourceData, кроме столбца BikeBuyer (это ведь новые клиенты, мы не знаем, сделают ли они покупку!), кроме того, есть ряд несущественных для анализа новых параметров - имя, адрес электронной почты, телефон и т.д.

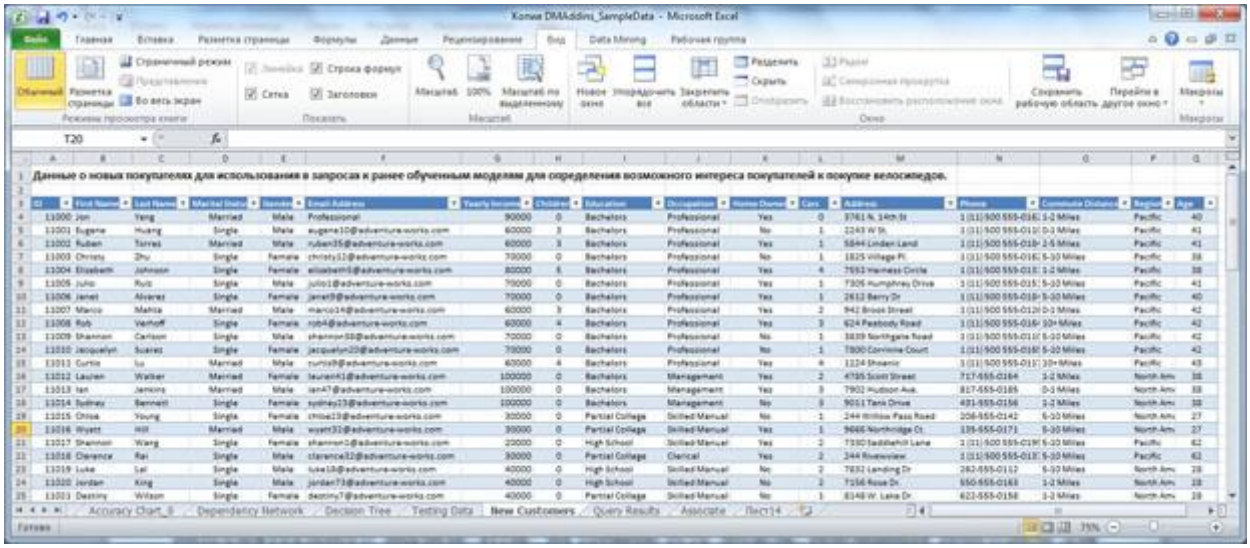
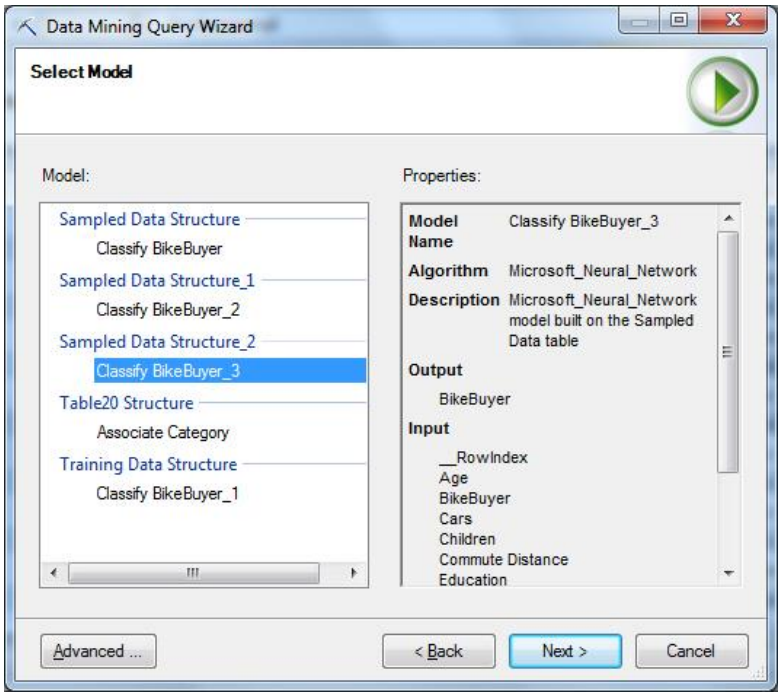


Рис. 15.9. Таблица NewCustomers

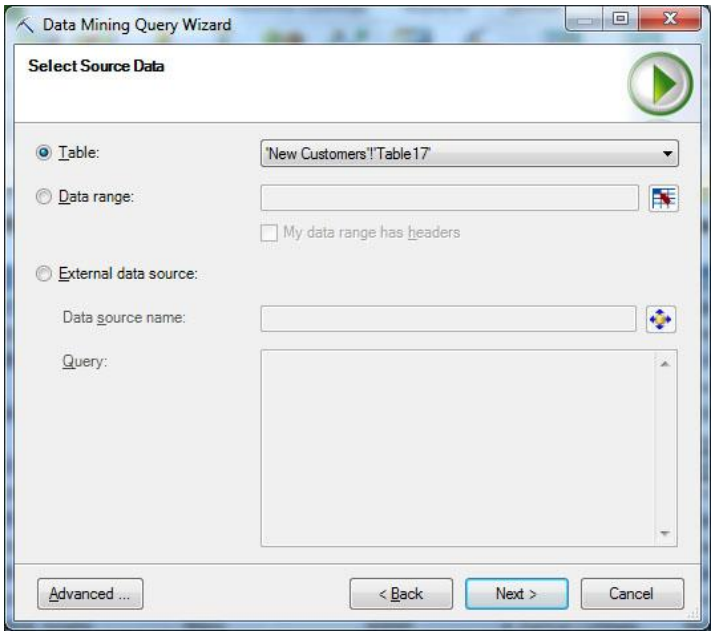
Наша задача заключается в том, чтобы предсказать, кто из этих людей готов сделать покупку. Запускаем инструмент Query (группа ModelUsage, (рис. 15.1) и выбираем используемую модель интеллектуального анализа (рис. 15.10-1). После этого указываем источник данных, для которого надо провести анализ. В нашем случае это таблица "NewCustomers" (рис. 15.10-2). Следующее окно позволяет указать соответствие параметров модели и столбцов таблицы. В нашем случае ничего исправлять не потребуется (рис. 15.10-3). Далее определяем выходное значение, т.е. столбец, который будет содержать прогноз. В окне "ChooseOutput" (аналогичном рис. 15.10-5, только без выходного значения), нажимаем кнопку "AddOutput" и получаем возможность определить выходной столбец (рис. 15.10-4). Назовем его "Будет покупать". В зависимости от того, куда будет выводиться результат работы (в исходную таблицу, на новый лист Excelи т.д.), может понадобиться включить в выходной набор дополнительные столбцы (идентификатор клиента и т.д.). После добавления выходных параметров (рис. 15.10-5) надо указать, куда будет выводиться результат. По умолчанию (рис. 15.10-6) он

попадет в таблицу с исходными данными, но можно потребовать вывод на новый или уже существующий лист Excel.

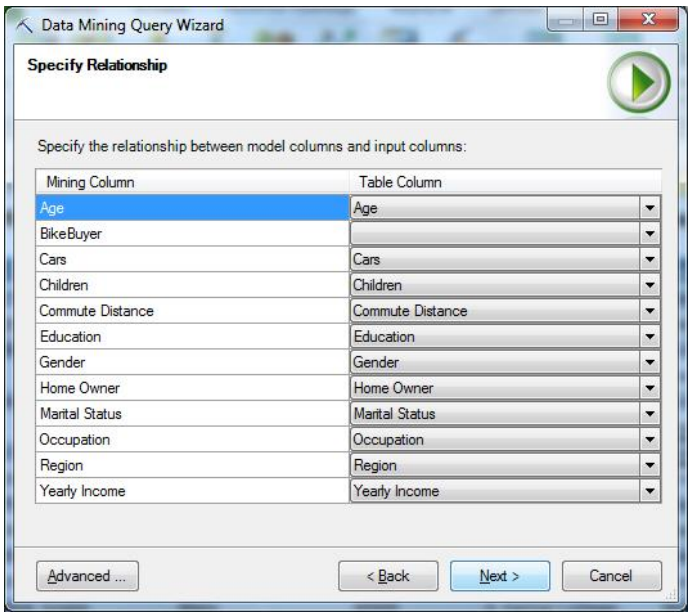
1.



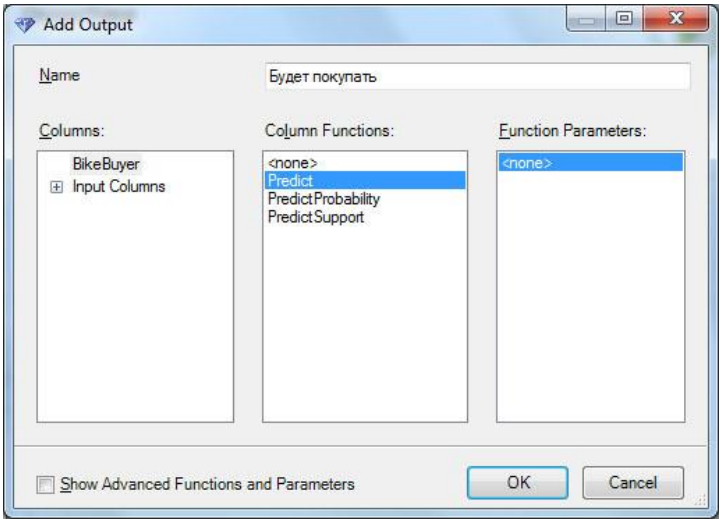
2.



3.


Data Mining Query Wizard
Specify Relationship
Specify the relationship between model columns and input columns:
Mining Column
Table Column
Age
Age
BikeBuyer
Cars
Cars
Children
Children
Commute Distance
Commute Distance
Education
Education
Gender
Gender
Home Owner
Home Owner
Marital Status
Marital Status
Occupation
Occupation
Region
Region
Yearly Income
Yearly Income
Advanced ...
< Back
Next >
Cancel

4.


Add Output
Name
Будет покупать
Columns:
Column Functions:
Function Parameters:
BikeBuyer
Input Columns
<none>
Predict
PredictProbability
PredictSupport
<none>
Show Advanced Functions and Parameters
OK
Cancel

5.

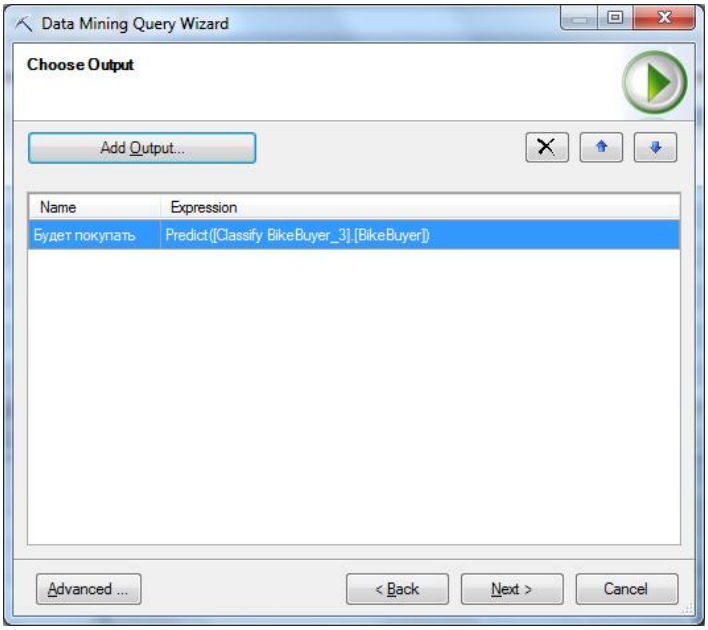
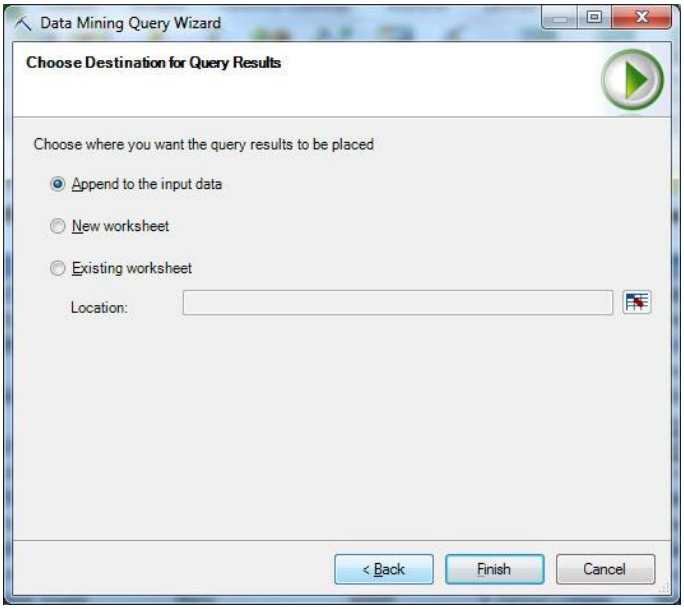


Рис. 15.10. Построение запроса

6.



В ходе работы в окнах рис. 15.10-1 - рис. 15.10-5 можно нажать кнопку Advanced и попасть в окно конструктора выражения на языке DMX (рис. 15.11). Здесь можно просмотреть или поправить генерируемый код запроса на DMX, который будет передан Аналитическим Службам MSSQLServer 2008.

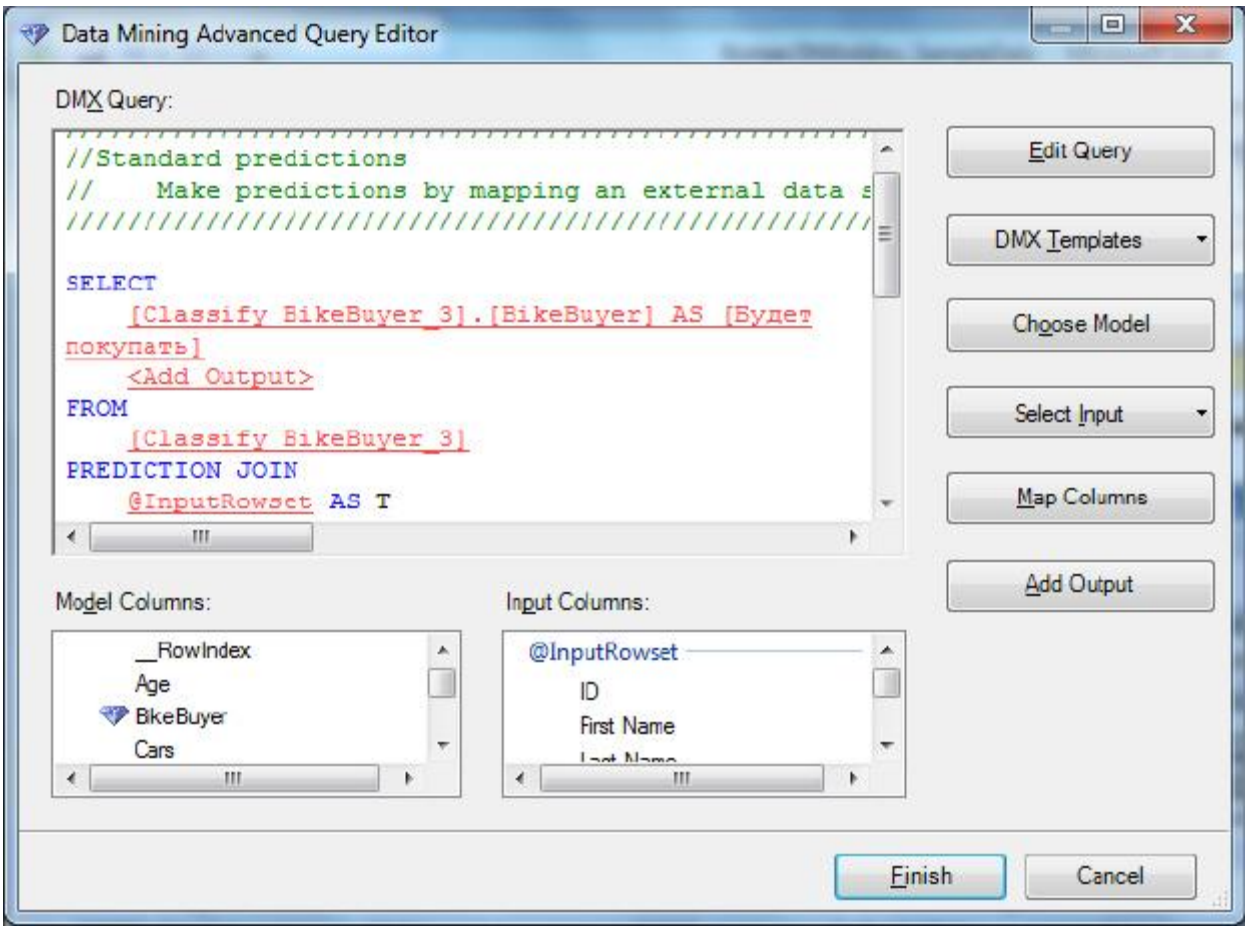


Рис. 15.11. Конструктор запросов

В результате выполнения сформированного мастером запроса в исходную таблицу будет добавлен столбец, содержащий результаты выполненной классификации (рис. 15.12).

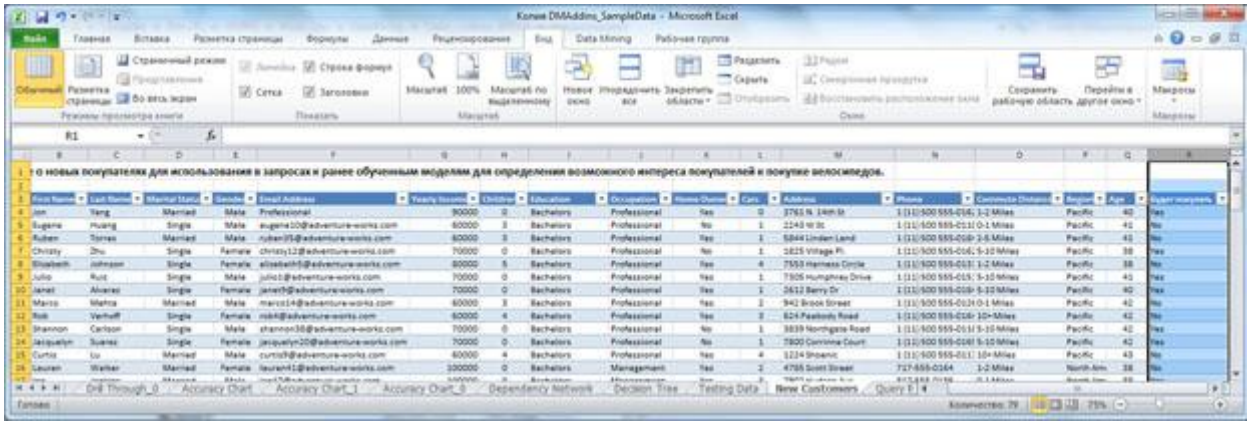


Рис. 15.12. В исходную таблицу добавлен столбец с результатами работы

Задание. Выполните запрос к модели интеллектуального анализа. Оцените полученные результаты.

Лабораторная работа 9. Построение модели кластеризации, трассировка и перекрестная проверка

Цель: В лабораторной работе рассматривается построение модели интеллектуального анализа данных, использующей алгоритм кластеризации, проводится анализ модели с использованием перекрестной проверки и рассматриваются предоставляемые DataMiningClient возможности по выполнению трассировки запросов к серверу.

Рассмотрим еще ряд возможностей, предоставляемых надстройками интеллектуального анализа данных.

Пусть необходимо провести сегментацию клиентов Интернет магазина, список которых находится в файле Excel. Если использовать TableAnalysisTools, для решения этой задачи надо применить инструмент DetectCategories(см. "Использование инструментов "AnalyzeKeyInfluencers" и "DetectCategories""). Также можно воспользоваться средствами DataMiningClientforExcel, где выбрать инструмент Cluster (рис. 16.1).

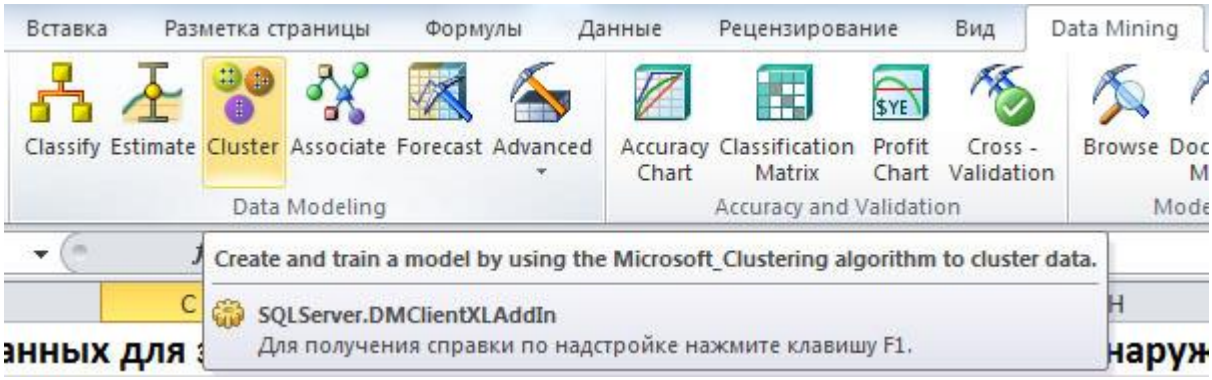


Рис. 16.1. Инструмент Cluster

Итак, откроем файл с образцами данных, идущий с надстройками интеллектуального анализа, перейдем на лист TableAnalysisToolsSample (или можно с первого листа с оглавлением перейти по ссылке "Образцы данных для средств анализа таблиц") и запустим инструмент Cluster.

Первое окно кратко описывает суть задачи кластеризации и указывает на то, что для работы мастера необходимо подключение к MS SQLServer (которое у нас было настроено ранее). Следующее окно (рис. 16.2-1) позволяет указать источник данных - в нашем случае это электронная таблица Excel, после чего можно выбрать число кластеров (рис. 16.2-2) или указать автоматическое определение, а также используемые столбцы входных данных. Здесь сбросим флажки рядом со столбцами ID и PurchasedBike.

1.

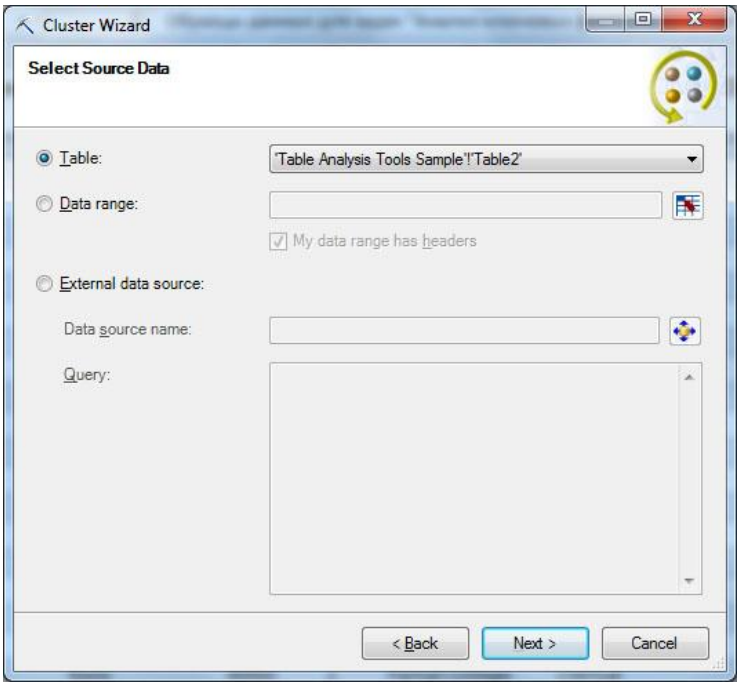
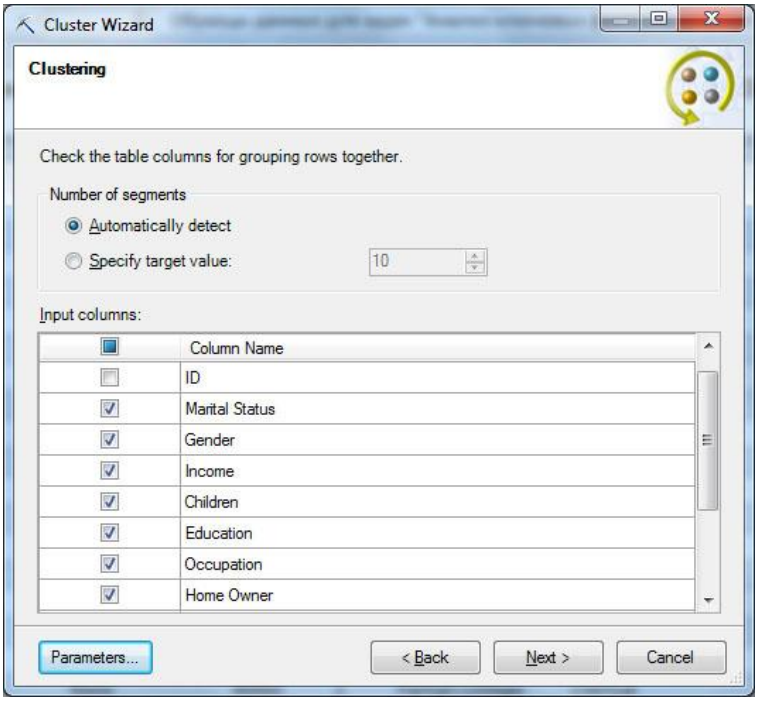


Рис. 16.2. Диалоговые окна мастера кластеризации

2.



Описанный выше выбор входных параметров обусловлен тем, что столбец с уникальным идентификатором покупателя может только помешать алгоритму кластеризации, а купил ли клиент велосипед или нет, нас сейчас не интересует. Кроме того, нажав в этом окне кнопку Parameters… можно получить доступ к настройке параметров алгоритма кластеризации (рис. 16.3) и, например, поменять используемый по умолчанию метод кластеризации.

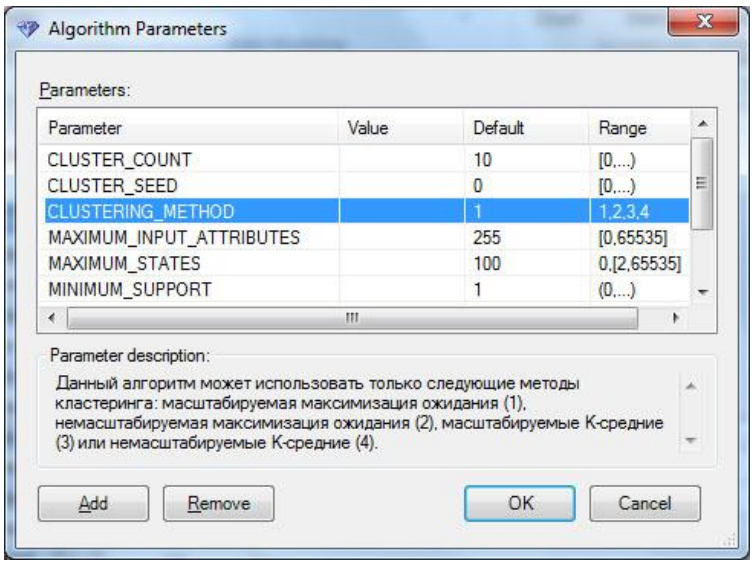


Рис. 16.3. Параметры модели кластеризации

Более подробно настройки алгоритма кластеризации обсуждаются в теоретической части курса. Следующее окно мастера позволяет указать процент данных, резервируемых для задач тестирования. И наконец, в последнем окне мастера (рис. 16.4) можно задать имя структуры и модели, указать, открывать ли просмотр модели, разрешить ли детализацию, использовать ли временные модели (по умолчанию - нет).

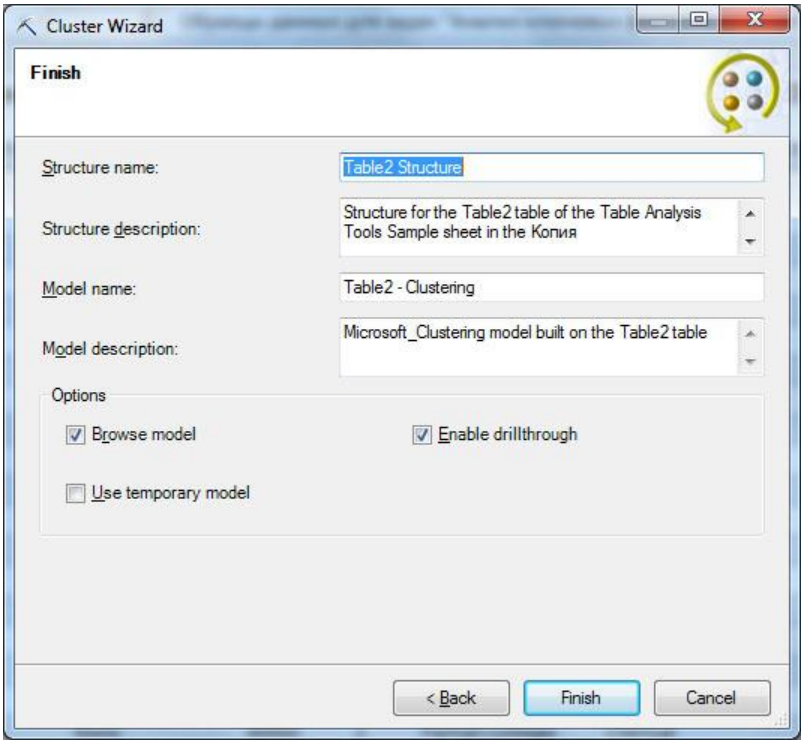


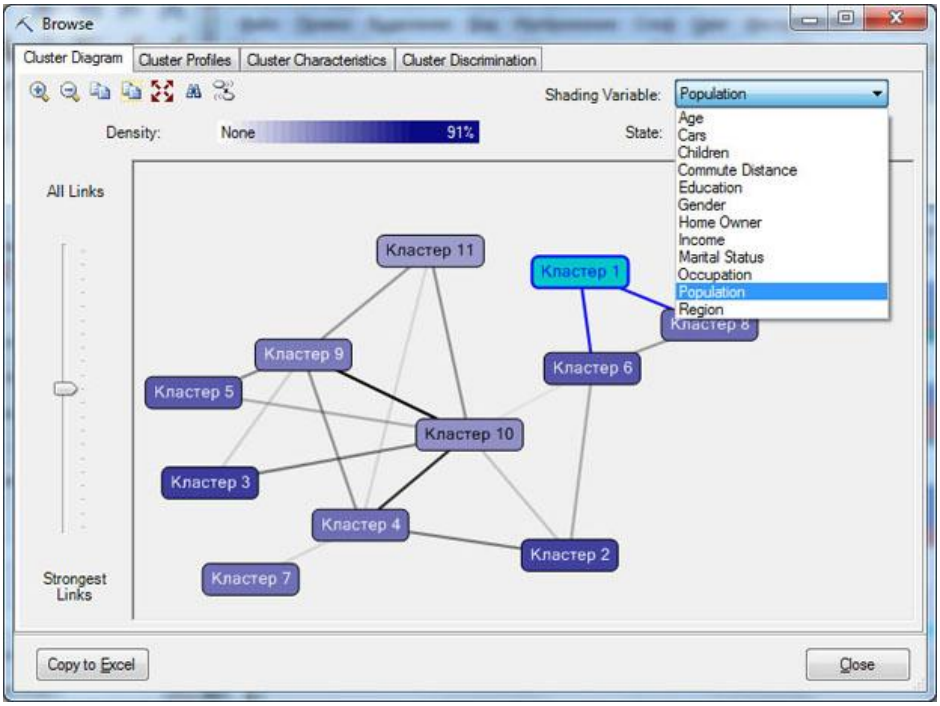
Рис. 16.4. Определение имен структуры и модели

После нажатия кнопки Finish будет создана структура и модель, после чего модель будет обработана и открыта для просмотра в окне Browser (рис. 16.5-1 - рис. 16.5-4). Диаграмма кластеров (рис. 16.5-1) отображает все

кластеры в модели, в нашем примере их 11. Заливка линии, соединяющей кластеры, показывает степень их сходства. Светлая или отсутствующая заливка означает, что кластеры не очень схожи. Можно выбрать анализ по отдельному атрибуту или по всей совокупности (Population).

Нажав кнопку CopytoExcel можно получить изображение на отдельный лист таблицы Excel.

1.



2.

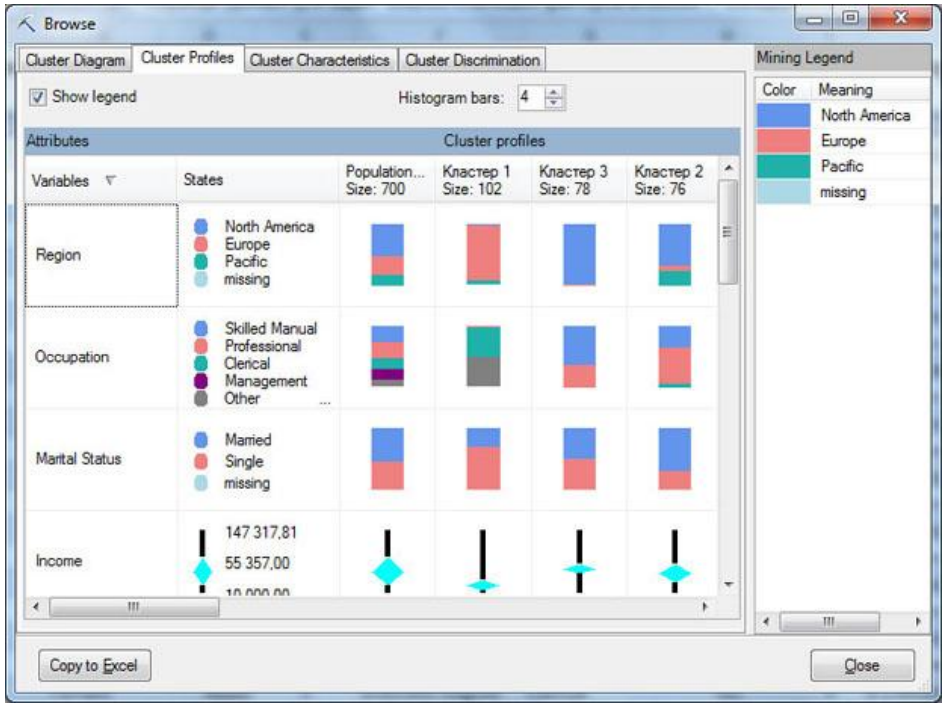


Рис. 16.5.

3.

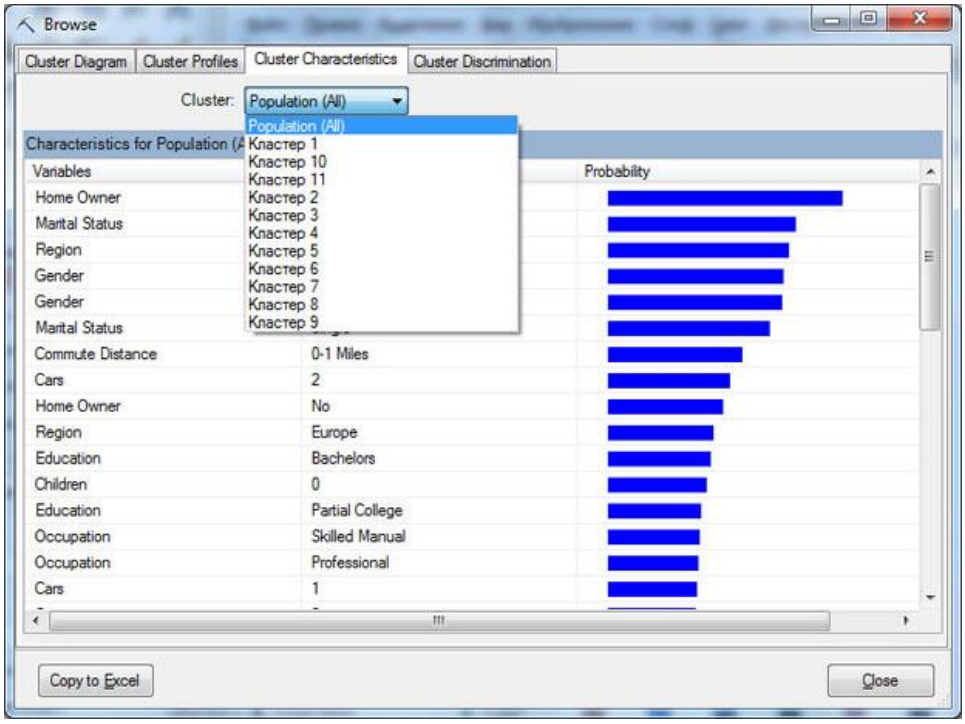
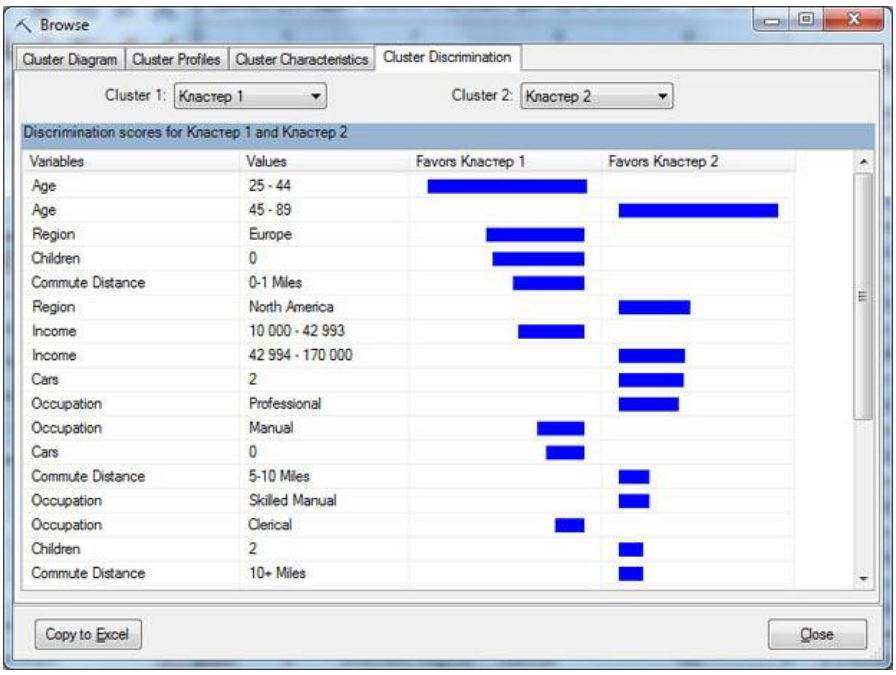


Рис. 16.5. Окна ModelBrowser

4.



Окно ClusterProfile позволяет просмотреть распределение значений атрибутов в каждом кластере. Например, на рис. 16.5-1 видно, что большая часть клиентов, отнесенных к кластеру 1, проживают в регионе Europe, а большинство из кластера 3 относится к региону NorthAmerica. Дискретные атрибуты представлены в виде цветных линий, непрерывные атрибуты в виде диаграммы ромбов, представляющей среднее значение и стандартное отклонение в каждом кластере. Параметр Histogram bars ("Столбцы

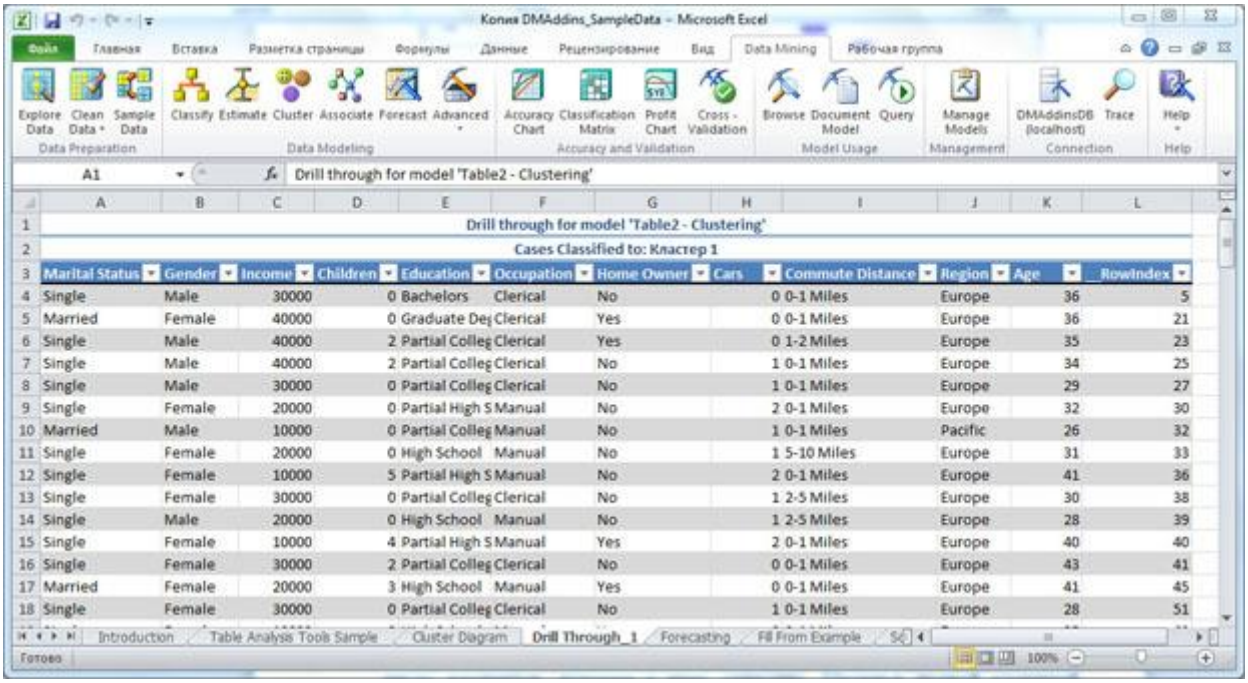
гистограммы") управляет количеством столбцов, видимых на гистограмме. Если доступно больше столбцов, чем выбрано для отображения, то наиболее важные столбцы сохраняются, а оставшиеся группируются в сегмент серого цвета.

В заголовке под названием каждого кластера указывает число вариантов, которые к нему отнесены. Щелкнув правой клавишей мыши на заголовке столбца, можно вызвать контекстное меню, позволяющее в частности переименовать соответствующий кластер. Кроме того, из контекстного меню, выбрав опцию DrillThroughModelColumn, можно получить детализацию модели (результаты выводятся на отдельный лист Excel). Например, на рис. 16.6 показаны все варианты, отнесенные к кластеру 1.

Но вернемся к окнам Modelbrowser. Окно ClusterCharacteristics позволяет просмотреть наиболее вероятные значения атрибутов для всего множества вариантов (Population) и для каждого кластера (если выбрать кластер в выпадающем списке). В последнем случае, столбцы сортируются по степени важности данного атрибута для кластера. Например, в рассмотренном выше кластере 1 на первом месте будет находиться атрибут Region со значением Europe. При этом, вероятность того что клиент, отнесенный алгоритмом к этой категории, проживает именно в Европе оценивается как очень высокая.

Окно ClusterDiscrimination позволяет провести попарное сравнение двух кластеров (рис. 16.5-4) или выбранного кластера и всех остальных вариантов.

Теперь перейдем к анализу того, что же происходит на сервере. В этом поможет инструмент Trace, расположенный в ленте DataMining в разделе Connection. Если нажать данную кнопку, откроется окно, в котором отображается содержимое отправляемых на сервер запросов (рис. 16.7).



Marital Status	Gender	Income	Children	Education	Occupation	Home Owner	Cars	Commute Distance	Region	Age	RowIndex
Single	Male	30000	0	Bachelors	Clerical	No	0	0-1 Miles	Europe	36	5
Married	Female	40000	0	Graduate De	Clerical	Yes	0	0-1 Miles	Europe	36	21
Single	Male	40000	2	Partial Colleg	Clerical	Yes	0	1-2 Miles	Europe	35	23
Single	Male	40000	2	Partial Colleg	Clerical	No	1	0-1 Miles	Europe	34	25
Single	Male	30000	0	Partial Colleg	Clerical	No	1	0-1 Miles	Europe	29	27
Single	Female	20000	0	Partial High S	Manual	No	2	0-1 Miles	Europe	32	30
Married	Male	10000	0	Partial Colleg	Manual	No	1	0-1 Miles	Pacific	26	32
Single	Female	20000	0	High School	Manual	No	1	5-10 Miles	Europe	31	33
Single	Female	10000	5	Partial High S	Manual	No	2	0-1 Miles	Europe	41	36
Single	Female	30000	0	Partial Colleg	Clerical	No	1	2-5 Miles	Europe	30	38
Single	Male	20000	0	High School	Manual	No	1	2-5 Miles	Europe	28	39
Single	Female	10000	4	Partial High S	Manual	Yes	2	0-1 Miles	Europe	40	40
Single	Female	30000	2	Partial Colleg	Clerical	No	0	0-1 Miles	Europe	43	41
Married	Female	20000	3	High School	Manual	Yes	0	0-1 Miles	Europe	41	45
Single	Female	30000	0	Partial Colleg	Clerical	No	1	0-1 Miles	Europe	28	51

Рис. 16.6. Результаты детализации модели

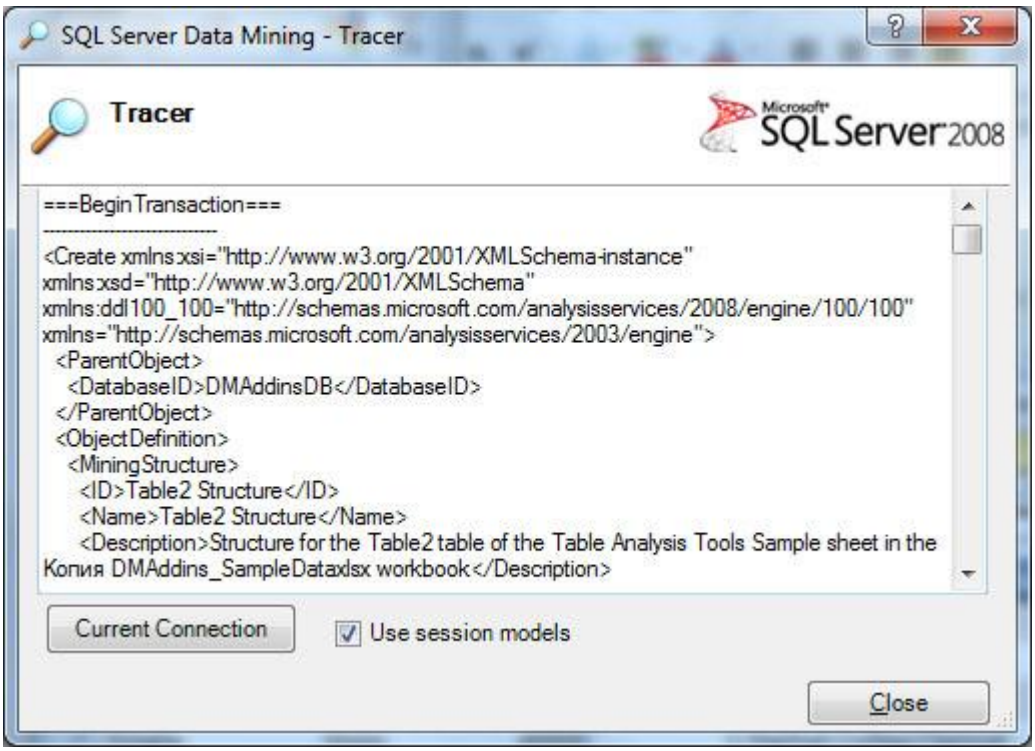


Рис. 16.7. Окно трассировки

Если проанализировать текст запросов, видно, что первая часть транзакции - это описание на XML создаваемой структуры и модели, вторая часть, которая приводится ниже - это DMX запрос на заполнение структуры (и, соответственно, обработку модели).

```
INSERT INTO MINING STRUCTURE [Table2 Structure] (__RowIndex,
    [Marital Status],
    [Gender],
    [Income],
    [Children],
    [Education],
    [Occupation],
    [Home Owner],
    [Cars],
    [Commute Distance],
    [Region],
    [Age]) @ParamTable
ParamTable =
Microsoft.SqlServer.DataMining.Office.Excel.ExcelDataReader
```

Листинг 16.1. DMX-запрос на обработку структуры

Использование трассировки позволяет глубже разобраться в особенностях работы надстроек интеллектуального анализа и при возникновении ошибок выявить их причины.

Задание 1. По аналогии с рассмотренным примером создайте модель кластеризации. Изучите и проанализируйте полученные результаты. Отройте окно трассировки, проанализируйте отправляемые на сервер запросы.

Теперь рассмотрим инструмент перекрестной проверки Cross-Validation (надо отметить, что перекрестная проверка доступна при использовании SQLServer версии Enterprise или Developer). Суть ее заключается в том, что множество вариантов, которые использует модель, разбивается на непересекающиеся подмножества (разделы), для каждого из которых производится обработка модели и полученные результаты сравниваются с теми, что были на исходном множестве вариантов. Если результаты близки, можно говорить об удачной модели интеллектуального анализа (исходных данных хватило, результат анализа/прогноза достаточно стабилен).

В разделе Accuracy and Validation выберем инструмент Cross-Validation. Первое окно мастера сообщает о сути выполняемой проверки. Во втором окне (рис. 16.8) производится выбор модели для перекрестной проверки. Укажем нашу модель кластеризации - Table2-Clustering.

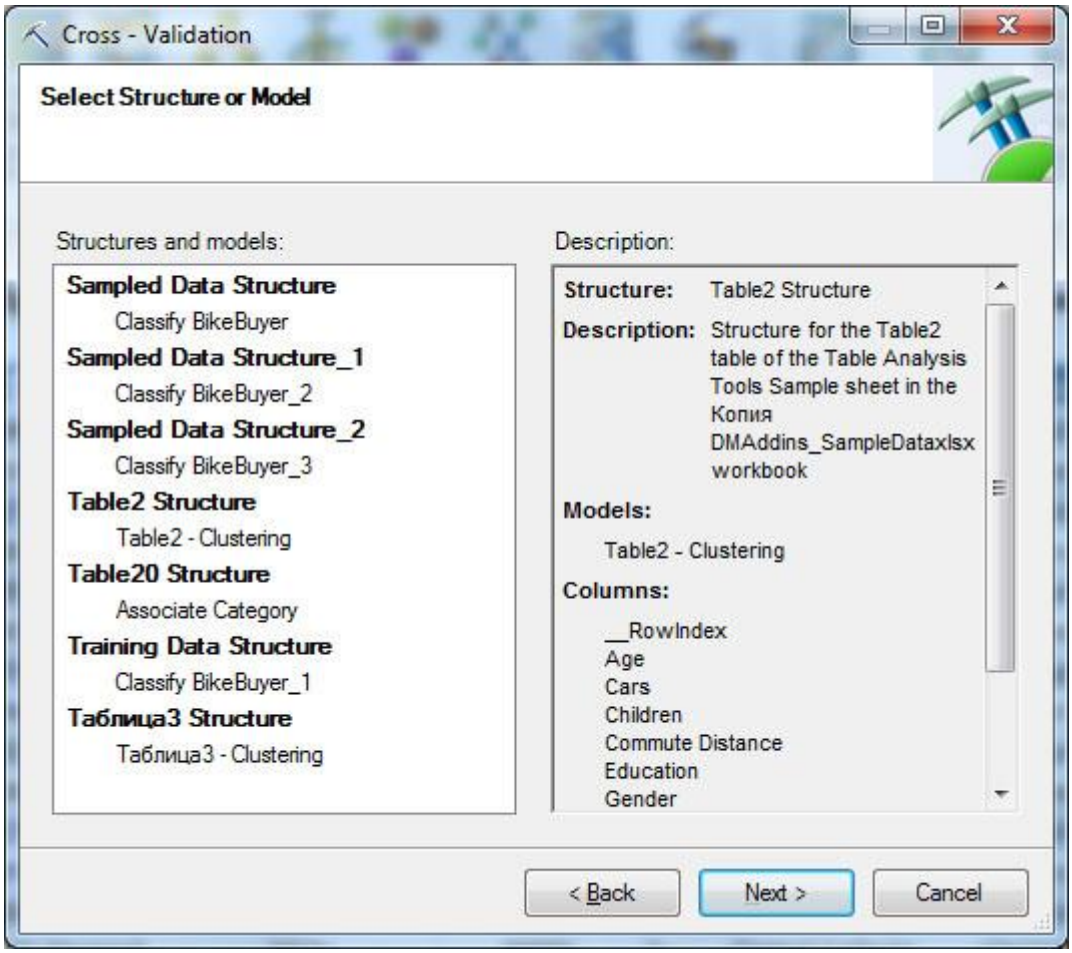


Рис. 16.8. Выбор модели для перекрестной проверки

После выбора модели нужно указать параметры проводимой перекрестной проверки. В частности, указывается число разделов с данными для перекрестной проверки (FoldCount, по умолчанию 10), максимальное число вариантов, используемых при проверке (значение MaximumRows=0 указывает на то, что будут использоваться все; если исходных данных много, при использовании всех данных перекрестная проверка может занять

продолжительное время), целевой атрибут (TargetAttribute). На рисунке стоит TargetAttribute#Cluster, т.е. номер кластера, к которому принадлежит вариант. Суть проверки будет заключаться в том, что выполняется кластеризация в рамках отдельного раздела и полученный номер кластера, к которому отнесен вариант, будет сравниваться с номером кластера, полученным при обработке модели с использованием всего множества вариантов. Совпадение говорит о том, что модель хорошая (правильно определены имеющиеся шаблоны).

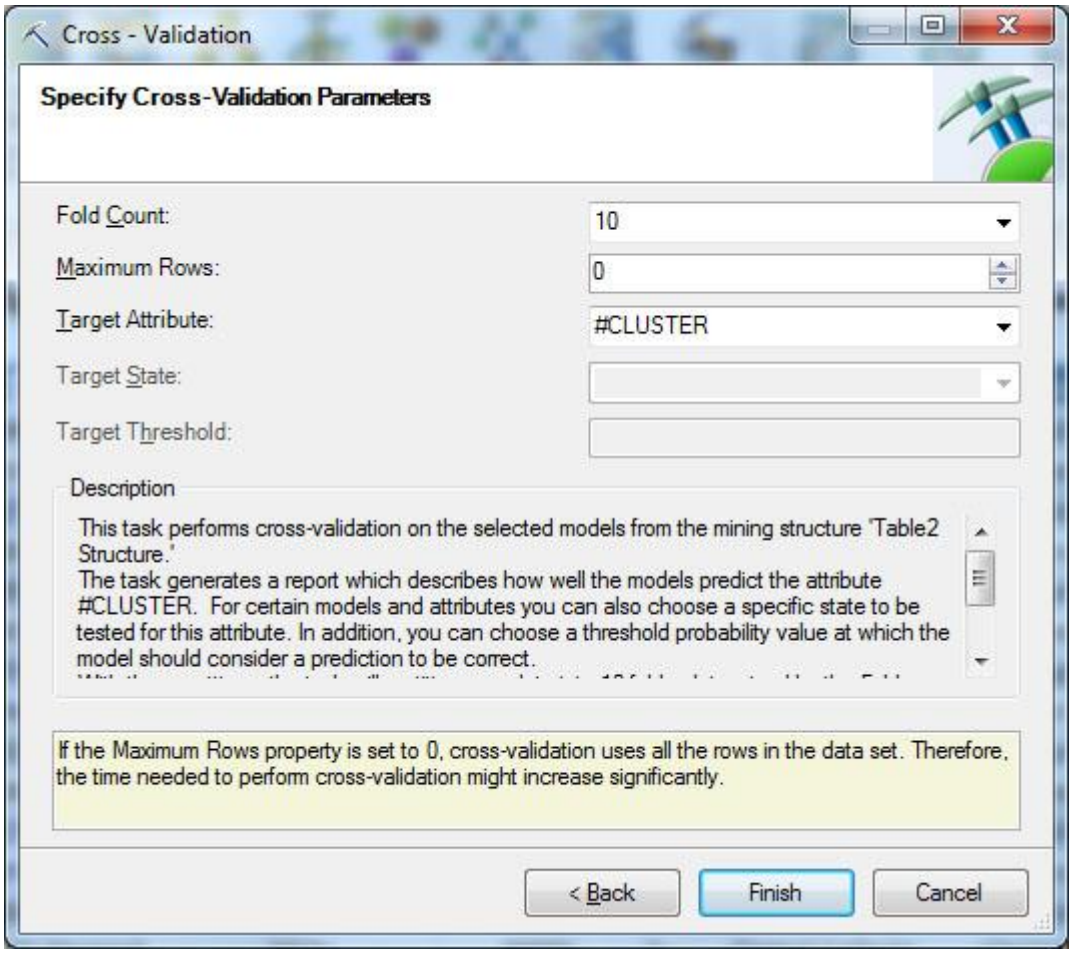


Рис. 16.9. Указание параметров перекрестной проверки

По результатам выполнения перекрестной проверки формируется отчет (рис. 16.10). В нем показывается, сколько вариантов использовалось для проверки (на рисунке - 700), какие разделы были сформированы (в нашем примере 10 разделов по 70 строк данных), результаты проведенного анализа. Отчет (рис. 16.10) показал, что в среднем, результаты, полученные при анализе по разделам, более чем в 82% случаев совпадают с результатами исходной модели. Небольшой разброс значений для разных разделов, указывает на стабильность получаемого результата, т.е. построенная модель интеллектуального анализа может быть признана удачной.

Задание 2. Выполните перекрестную проверку для созданной модели интеллектуального анализа. Опишите и проанализируйте полученные

результаты.

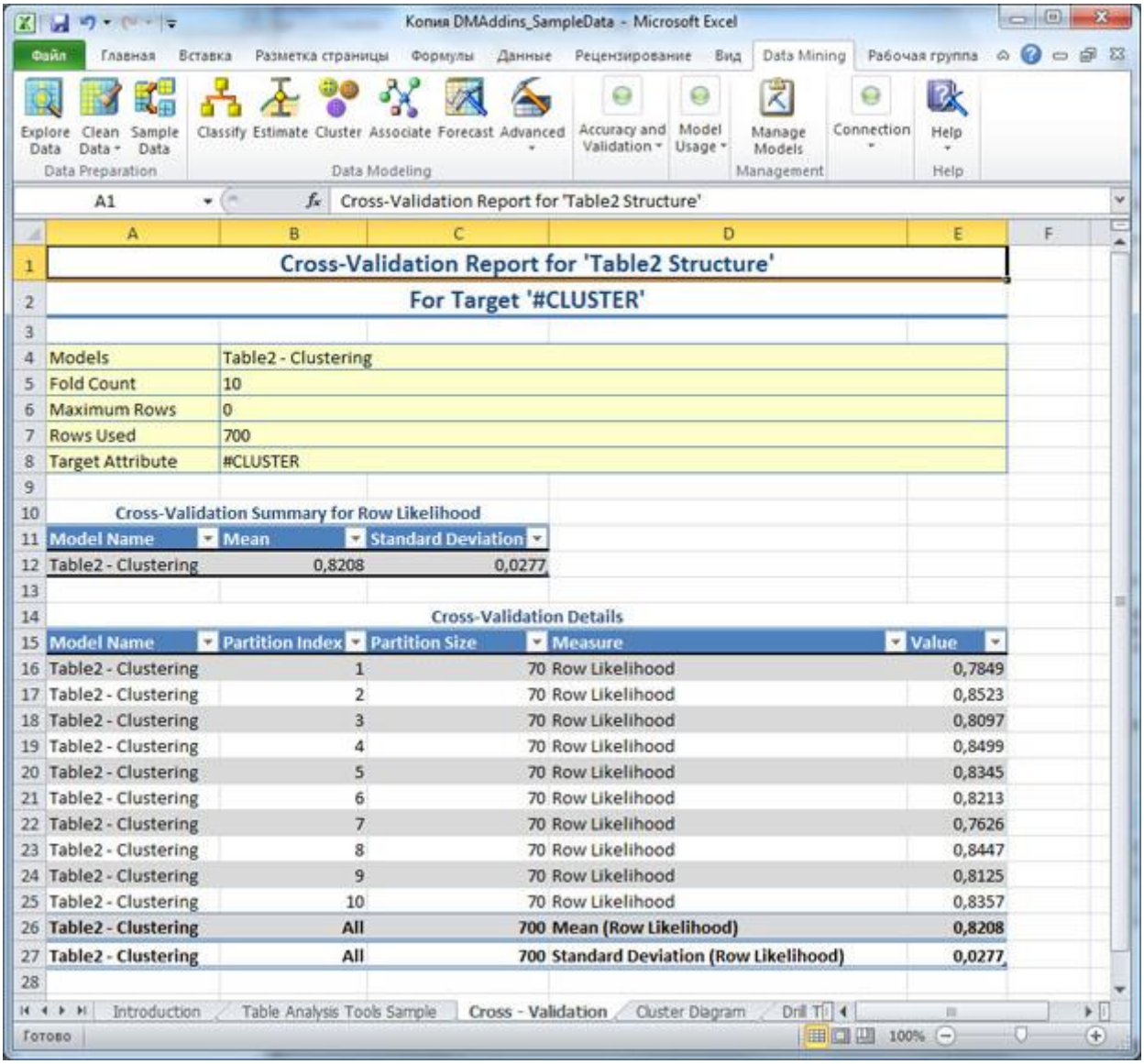


Рис. 16.10. Отчет по результатам проверки