ФЕДЕРАЛЬНОЕ АГЕНТСТВО СВЯЗИ

Федеральное государственное образовательное бюджетное учреждение высшего профессионального образования
«САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ТЕЛЕКОММУНИКАЦИЙ им. проф. М. А. БОНЧ-БРУЕВИЧА»

М. Б. Вольфсон

АНАЛИЗ ДАННЫХ

МЕТОДИЧЕСКИЕ УКАЗАНИЯ
К ЛАБОРАТОРНЫМ И ПРАКТИЧЕСКИМ РАБОТАМ

СПбГУТ)))

САНКТ-ПЕТЕРБУРГ
2018

ЛАБОРАТОРНАЯ РАБОТА №1. Анализ данных в реляционной модели

Исходные данные

Имеется компания, занимающаяся продажей продуктов питания. Для хранения и последующего анализа выбираются данные, описывающие бизнес-процесс продажи товаров оптовым клиентам.

Объектами предметной области являются клиенты, товары, сотрудники отдела продаж, поставщики.

Задание

1. В Microsoft Access 2007 создать реляционную базу данных (рис. 1.1). Имя файла должно соответствовать фамилии студента.

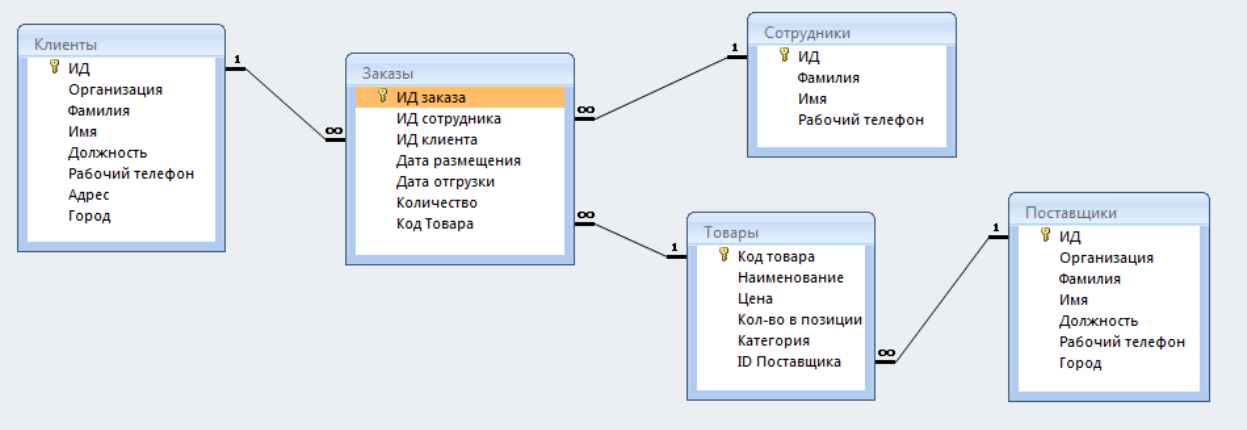


Рис. 1.1

Для более быстрого заполнения таблиц рекомендуется создать маски ввода и использовать операции копирования и вставки.

Таблица *Заказы*, не менее 30 записей (рис. 1.2).

ИД заказа	Сотрудник	Клиент	Дата размещения	Дата отгрузки	Количество	Код Товара
30	9	1	15.01.2010	22.01.2010	100	NWTB-1
31	3	2	20.01.2010	22.01.2010	200	NWTB-10
32	4	3	22.01.2010	22.01.2010	300	NWTB-11
33	6	4	30.01.2010	31.01.2010	400	NWTB-12
34	9	5	06.02.2010	07.02.2010	500	NWTB-13
35	3	6	10.02.2010	12.02.2010	600	NWTB-14
36	4	7	23.02.2010	25.02.2010	700	NWTB-15
37	8	8	06.03.2010	09.03.2010	800	NWTB-16
38	9	9	10.03.2010	11.03.2010	900	NWTB-17
39	3	10	22.03.2010	24.03.2010	100	NWTB-19
40	4	11	24.03.2010	24.03.2010	200	NWTB-20

Рис. 1.2

Таблица *Клиенты*, не менее 10 записей (рис. 1.3).

ИД	Организация	Фамилия	Имя	Рабочий телефон	Адрес	Должность	Город
1	Defa	Костерина	Ольга	(123)555-0111	1-я улица, д. 12	Ответственный	Сочи
2	ООО "Эдельвейс"	Верный	Григорий	(123)555-0112	2-я улица, д. 12	Ответственный	Иваново
3	ООО "Алкон"	Егоров	Владимир	(123) 555-0113	3-я улица, д. 12	Сотрудник отдела снабжения	Уфа
4	ООО "Альтаир"	Омельченко	Светлана	(123) 555-0114	4-я улица, д. 12	Начальник отдела снабжения	Москва
5	ОАО "Арго"	Песоцкий	Станислав	(123) 555-0115	5-я улица, д. 12	Ответственный	Иркутск
6	ООО "Балт-Конд"	Шашков	Руслан	(123) 555-0116	6-я улица, д. 12	Начальник отдела снабжения	Казань
7	ЗАО "Глобар"	Вронский	Юрий	(123) 555-0117	7-я улица, д. 12	Ответственный	Воронеж
8	ЗАО "ЛЭНД"	Подколзина	Екатерина	(123) 555-0118	8-я улица, д. 12	Сотрудник отдела снабжения	Омск
9	ЗАО "Камея"	Ерёменко	Алексей	(123) 555-0119	9-я улица, д. 12	Начальник отдела снабжения	Пермь
10	ООО "Клин"	Грачев	Николай	(123) 555-0120	10-я улица, д. 12	Начальник отдела снабжения	Саратов
11	ООО "Локис"	Орехов	Алексей	(123) 555-0121	11-я улица, д. 12	Начальник отдела снабжения	Орел
12	ООО "Риф"	Володин	Виктор	(123) 555-0122	11-я улица, д. 12	Начальник отдела снабжения	Тюмень
13	ЧП "Синицын"	Туманов	Александр	(123) 555-0123	13-я улица, д.4	Сотрудник отдела снабжения	Курск

Рис. 1.3

Таблица *Сотрудники*, не менее 10 записей (рис. 1.4).

ИД	Имя	Рабочий телефон	Фамилия
1	Юлия	(123)555-0101	Ильина
2	Андрей	(123) 555-0102	Гладких
3	Евгений	(123) 555-0103	Куликов
4	Мария	(123) 555-0104	Сергиенко
5	Николай	(123) 555-0105	Новиков
6	Вадим	(123) 555-0106	Корепин
7	Сергей	(123) 555-0107	Климов
8	Инна	(123) 555-0108	Ожогина
9	Дарья	(123) 555-0109	Попкова

Рис. 1.4

Таблица *Поставщики*, не менее 10 записей (рис. 1.5).

ИД	Организация	Фамилия	Имя	Рабочий телефон	Город	Должность
1	ООО "Аврора"	Андреева	Елизавета	1234567	Санкт-Петербург	Начальник отдела сбыта
2	ООО "Балтика"	Волкова	Марина	2345678	Санкт-Петербург	Начальник отдела сбыта
3	ИП "Волга"	Котова	Маргарита	3456789	Москва	Сотрудник отдела сбыта
4	ИП "Горизонт"	Орлов	Николай	4567890	Москва	Маркетолог
5	ОАО "Диета"	Зорин	Антон	5678901	Москва	Начальник отдела сбыта
6	ОАО "Питер"	Хромов	Евгений	6789012	Санкт-Петербург	Помощник по маркетингу
7	ЗАО "Нева"	Гласнов	Олег	7890123	Санкт-Петербург	Маркетолог
8	ЗАО "Белые ночи"	Сидоров	Борис	8901234	Санкт-Петербург	Сотрудник отдела сбыта
9	ООО "Ладога"	Соколова	Лариса	9012345	Санкт-Петербург	Начальник отдела сбыта

Рис. 1.5

Таблица *Товары*, не менее 30 записей, не более 10 категорий (рис. 1.6).

Код товара	Наименование	Цена	Кол-во в позиции	Категория	ID Поставщика
NWTB-1	Ананас	1,80р.	15,25 унций	Фруктовые и овощные консервы	9
NWTB-10	Зеленый горошек	1,50р.	14,5 унций	Фруктовые и овощные консервы	4
NWTB-11	Зеленый чай	2,99р.	20 пакетиков в коробке	Напитки	1
NWTB-12	Индийский чай	4,00р.	100 штук в коробке	Напитки	6
NWTB-13	Карри	40,00р.	12 банок по 400 г	Соусы	7
NWTB-14	Конфеты	11,80р.	5 коробок	Кондитерские изделия	3
NWTB-15	Копченый лосось	4,00р.	5 унций	Мясные консервы	6
NWTB-16	Кофе	46,00р.	16 банок по 500 г	Напитки	7
NWTB-17	Кукуруза	1,20р.	14,5 унций	Фруктовые и овощные консервы	3
NWTB-19	Лаваш	10,00р.	24 пакета по 4 штуки	Хлебобулочные изделия	3
NWTB-20	Луизианский соус	21,05р.	32 бутылки по 8 унций	Соусы	4
NWTB-21	Мармелад	81,00р.	30 коробок	Варенье и джемы	2
NWTB-22	Миндаль	10,00р.	пакет 5 кг	Сухофрукты и орехи	7
NWTB-28	Персики	1,50р.	15,25 унций	Фруктовые и овощные консервы	8
NWTB-29	Пиво	14,00р.	24 бутылки по 12 унций	Напитки	4
NWTB-3	Вишневый пирог	2,00р.	15,25 унций	Фруктовые и овощные консервы	1
NWTB-30	Пирожное с орехами	12,49р.	3 коробки	Хлебобулочные изделия	4

Рис. 1.6

2. В созданной базе данных создать запросы на выборку данных (при необходимости осуществив группировку данных) и проанализировать результаты.

2.1. Вывести фамилию самого успешного сотрудника по критерию количества продаж (*Запрос 2_1*)

2.2. Вывести фамилию самого успешного сотрудника по критерию объема выручки (*Запрос 2_2*).

2.3. Вывести самый продаваемый товар по количеству заказов (Запрос 2_3).

2.4. Вывести поставщика, продукция которого оказалась самой востребованной (*Запрос 2_4*).

2.5. Вывести самого выгодного клиента по критерию количества закупаемого товара (*Запрос 2_5*).

2.6. Вывести средний объем заказа в количественном и денежном выражении для каждого клиента (*Запрос 2_6*).

2.7. Вывести минимальное и максимальное количество дней между размещением и отгрузкой заказов (*Запрос 2_7*).

3. Эмулируя создание многомерной таблицы создать перекрестные запросы (использовать Мастер запросов) и проанализировать полученные результаты.

Для вывода полей из разных таблиц предварительно создайте запрос (*Запрос_темп*), содержащий все необходимые для запросов п.3 поля. Обратите внимание на многократное повторение данных в полученной таблице.

3.1. Вывести распределение принятых заказов каждого сотрудника по месяцам и определить количество заказов за каждый месяц (*Запрос 3_1*).

3.2. Вывести количество купленных товаров каждым из клиентов по товарным категориям и выявить самую популярную категорию (*Запрос 3_2*).

3.3. Вывести распределение количества поставок по товарным категориям и городам фирм-поставщиков (*Запрос 3_3*).

ЛАБОРАТОРНАЯ РАБОТА №2. Анализ данных в многомерной модели

Целью лабораторной работы является изучение и анализ возможностей реляционной модели хранения данных для последующего бизнес-анализа.

Исходные данные

Имеется компания, занимающаяся продажей продуктов питания. Для хранения и последующего анализа выбираются данные, описывающие бизнес-процесс продажи товаров оптовым клиентам.

Объектами предметной области являются клиенты, товары, сотрудники отдела продаж, поставщики.

Задание

1. Создать новую книгу Excel. Имя файла должно соответствовать фамилии студента.

2. Первый лист назвать «*Справочники*». Выбрав пункт меню «Получить данные из Access» последовательно импортировать на этот лист содержание таблиц «*Товары*», «*Клиенты*», «*Поставщики*», «*Сотрудники*», снабдив таблицы соответствующими заголовками.

Таким образом, мы получаем следующие ключевые измерения данных: *название* и *цена товара, организация покупателя, фамилия сотрудника и организация поставщика*. Кроме того, имеются несколько дополнительных измерений данных: *категория продукта, город поставщика* и др.

3. Второй лист назвать «*Заказы*». Импортировать на этот лист содержание таблицы «*Заказы*» (при желании, впоследствии подключения к базе данных можно удалить через меню «*Данные-Подключения*»).

4. Заменить внешние ключи реляционной базы даны (*коды сотрудников, клиентов и товаров*) на названия, взятые из таблиц листа «*Справочник*». Для этого создать новые поля и воспользоваться функцией ПРОСМОТР (кодовые поля на листе «*Справочник*» должны быть отсортированы по возрастанию).

5. Аналогично п. 4. добавить в таблицу на лист «*Заказы*» измерения *Категория товара, Цена, Название поставщика* и *Город поставщика*. Кодовые поля на листе «*Заказы*» можно скрыть.

Таким образом, мы получаем исходную таблицу для последующего многомерного анализа, имеющую измерения *Сотрудник, Клиент, Товар, Категория, Цена, Поставщик, Город* и факты *Дата размещения, Дата отгрузки, Количество*.

6. Построить сводную таблицу, соответствующую кубу данных с измерениями *Клиент, Товар, Месяц размещения заказа* и с набором фактов о выручке с продаж. В качестве дополнительных измерений (фильтров сводной таблицы) выбрать *Категорию товара, Поставщика* и *Сотрудника*.

7. Оформить таблицу по образцу рис. 2.1.

8. Проанализировать полученный результат и получить ответы на следующие вопросы (результаты сохранить на отдельных листах Excel).

8.1. Каков объем выручки за зимние месяцы от продажи фруктов и продуктов из фруктов?

8.2. Каков итоговый объем выручки от продажи товаров поставщиков – акционерных обществ, реализованных первыми тремя сотрудниками?

	A	B	C	D	E	F	G	H
1	Категория	(Все)						
2	Поставщик	(Все)						
3	Сотрудник	(Все)						
4								
5	Сумма по полю Оборот	Названия столбцов						
6	Названия строк	янв	фев	мар	апр	май	июн	Общий итог
7	Defa							
8	Ананас	180,00р.	-	-	-	-	-	180,00р.
9	Карри	-	-	-	-	-	20 000,00р.	20 000,00р.
10	Персики	-	-	750,00р.	-	-	-	750,00р.
11	Фасоль	-	-	-	120,00р.	-	-	120,00р.
12	**Defa Итог**	**180,00р.**	**-**	**750,00р.**	**120,00р.**	**-**	**20 000,00р.**	**53 400,00р.**
13	ЗАО "Глобар"							
14	Грецкие орехи	-	-	-	-	16 275,00р.	-	16 275,00р.
15	Копченый лосось	-	2 800,00р.	-	-	-	-	2 800,00р.
16	Луизианский соус	-	-	-	-	-	4 210,00р.	4 210,00р.
17	Сушеные сливы	-	-	-	700,00р.	-	-	700,00р.
18	**ЗАО "Глобар" Итог**	**-**	**2 800,00р.**	**-**	**700,00р.**	**16 275,00р.**	**4 210,00р.**	**93 240,00р.**
19	ЗАО "Камея"							
20	Ежевичный джем	-	-	-	-	22 500,00р.	-	22 500,00р.
21	Кукуруза	-	-	1 080,00р.	-	-	-	1 080,00р.
22	Миндаль	-	-	-	4 000,00р.	-	-	4 000,00р.
23	Тихоокеанские крабы	-	-	-	7 360,00р.	-	-	7 360,00р.
24	**ЗАО "Камея" Итог**	**-**	**-**	**1 080,00р.**	**22 720,00р.**	**22 500,00р.**	**-**	**141 960,00р.**
25	ЗАО "ЛЭНД"							
26	Груши	-	-	-	-	1 040,00р.	-	1 040,00р.
27	Кофе	-	-	36 800,00р.	-	-	-	36 800,00р.
28	Мармелад	-	-	-	24 300,00р.	-	-	24 300,00р.
29	Сушеные яблоки	-	-	-	15 900,00р.	-	-	15 900,00р.
30	**ЗАО "ЛЭНД" Итог**	**-**	**-**	**36 800,00р.**	**80 400,00р.**	**1 040,00р.**	**-**	**398 860,00р.**
31	ОАО "Арго"							
32	Карри	-	20 000,00р.	-	-	-	-	20 000,00р.
33	Кукуруза	-	-	-	-	-	1 080,00р.	1 080,00р.

Рис. 2.1.

9. На отдельном листе построить новую сводную таблицу, показывающую распределение объема продаж по городам закупок по месяцам. На том же листе построить гистограмму, показывающую итоги продаж по городам (рис. 2.2).

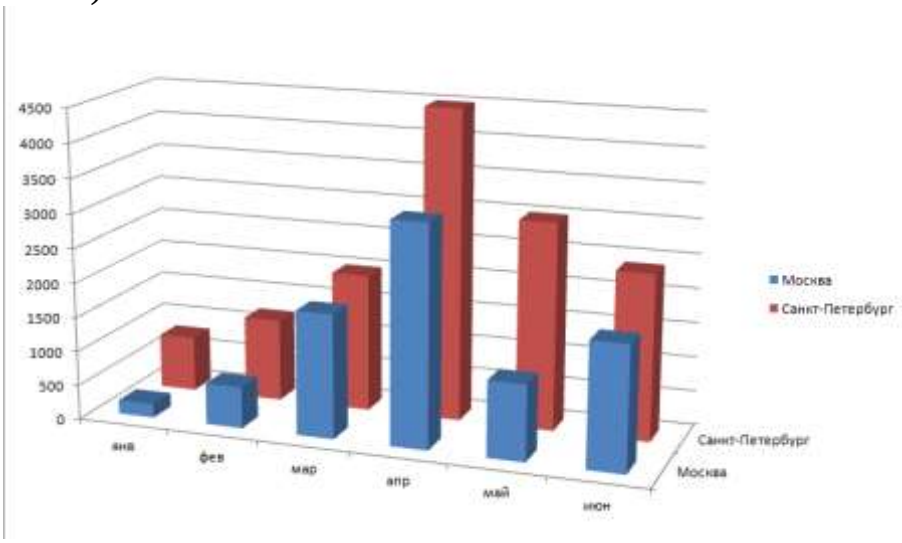


Рис. 2.2

ЛАБОРАТОРНАЯ РАБОТА №3. Работа с OLAP кубом

Исходные данные

Имеется компания, занимающаяся продажей продуктов питания. Для хранения и последующего анализа выбираются данные, описывающие бизнес-процесс продажи товаров.

Объектами предметной области являются продавцы и продукты.

Задание

1. В MS Excel создать новую книгу "*OLAP*".

2. В ту же папку сохранить с сайта www.fem-sut.spb.ru файл *nwdata_cube.cub*.

3. Через пункт меню «*Вставка-Сводная*» таблица запустить Мастер, выбрать пункт *'использовать внешний источник данных*', нажать кнопку '*выбрать подключение*' и указать в качестве источника файл *nwdata_cube.cub*.

4. В открывшемся справа окне изучить структуру OLAP куба.

5. Используя операцию **Сечения**, построить сводную таблицу с измерениями *Страна* (по столбцам) и *Год продажи* (по строкам). В качестве фактов взять количество проданных товаров. Агрегировать данные по странам и по годам.

6. Используя операцию **Транспонирования**, поменять местами измерения Страна и Год продажи.

7. Используя операцию **Сечения**, добавить в качестве измерения Категорию продукта.

8. Используя операцию **Детализации**, отобразить в таблице названия компаний и названия продуктов.

9. Используя операцию **Свертки**, скрыть названия продуктов по всей таблице.

10. Добавить агрегированное значение по сумме продаж по странам Аргентина и Австрия. Названную группу назвать «*Страны на А*». Установить показ промежуточных итогов в нижней части группы.

11. Сравнить полученный результат с рис. 3.1. Убедиться в невозможности изменять заданную в кубе иерархию уровней.

12. Скопировать полученную сводную таблицу на новые листы и построить на них диаграммы, позволяющие получить ответы на следующие вопросы (предварительно при необходимости внести изменения в структуру скопированных таблиц):

а) Доли пяти самых крупных стран продавцов по итоговым продажам (рис. 3.2.).

б) Объемы продаж самых продаваемых товаров из каждой товарной категории по годам (рис. 3.3.).

с) Динамика продаж морепродуктов в США по месяцам (рис. 3.4.).

	A	B	C	D	E
1	Sum Of Quantity	Названия столбцов			
2	Названия строк	1996	1997	1998	Общий итог
3	Страны на A				
4	Argentina				
5	Beverages		3	79	82,
6	Condiments		10	35	45,
7	Confections		29	28	57,
8	Dairy Products		3	51	54,
9	Grains/Cereals			20	20,
10	Produce		19	14	33,
11	Seafood		30	18	48,
12	Argentina Итог		94	245	339,
13	Austria				
14	Beverages	188	335	459	982,
15	Condiments	184	410	126	720,
16	Confections	65	393	117	575,
17	Dairy Products	212	430	385	1027,
18	Grains/Cereals	60	220	300	580,
19	Meat/Poultry	14	283	65	362,
20	Produce	99	201	88	388,
21	Seafood	127	75	331	533,
22	Austria Итог	949	2347	1871	5167,
23	Страны на A Итог	949,	2441,	2116,	5506,
24	Другие				
25	Belgium				
26	Beverages	12	92	168	272,
27	Condiments		60	87	147,
28	Confections	40	123	107	270,
29	Dairy Products	65	110	120	295,
30	Grains/Cereals		67	78	145,
31	Meat/Poultry	40	14	35	89,
32	Produce	28	50	20	98,
33	Seafood			76	76,

Рис. 3.1

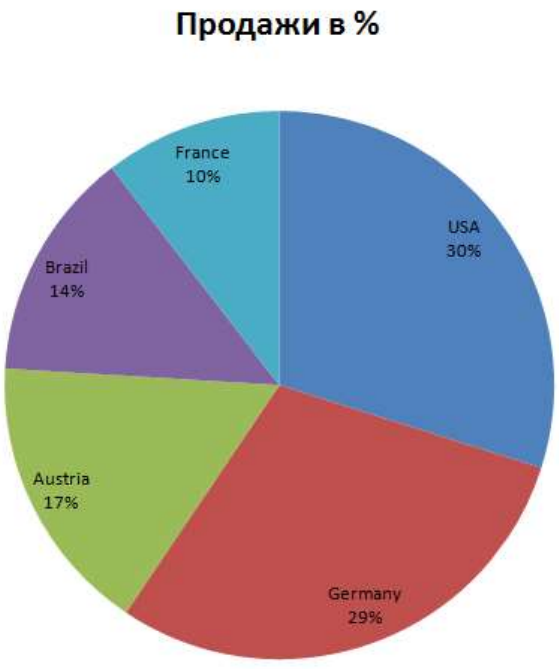


Рис. 3.2

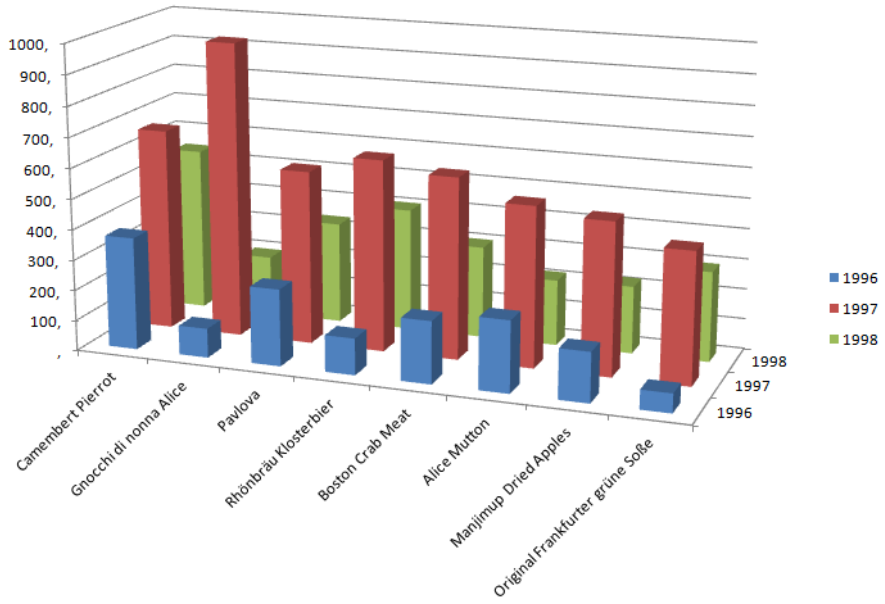


Рис. 3.3

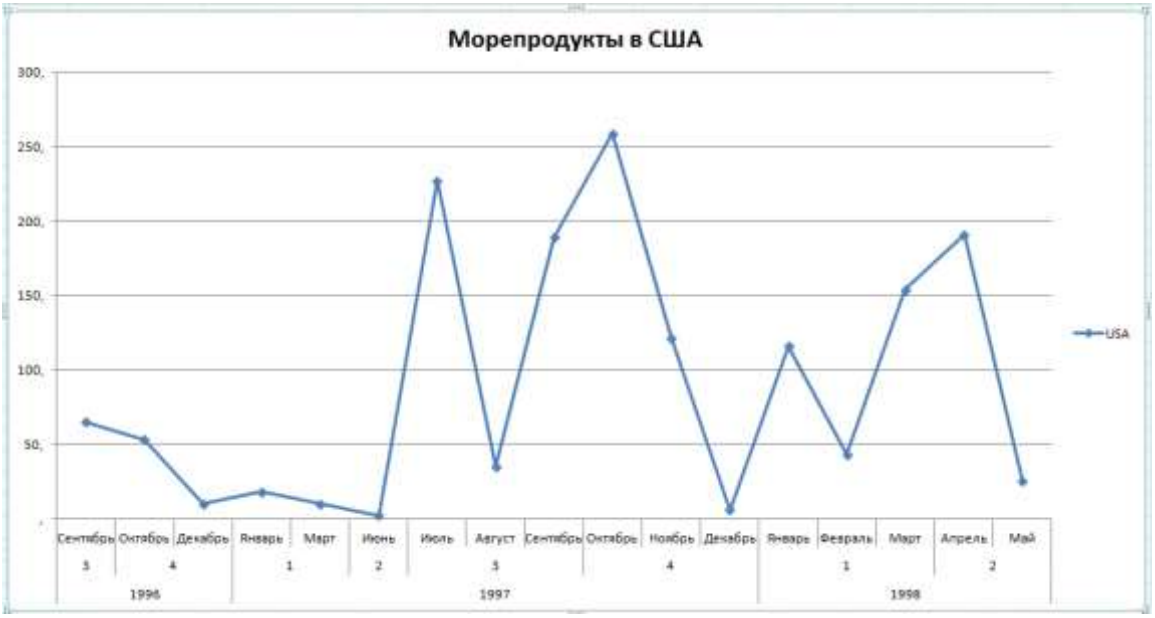


Рис. 3.4

ЛАБОРАТОРНАЯ РАБОТА №4. Знакомство с аналитической платформой Deductor. Предобработка данных.

Краткая характеристика пакета Deductor

Deductor Studio – программа, являющаяся составной частью платформы Deductor. Она содержит механизмы импорта, обработки, визуализации и экспорта данных для быстрого и эффективного анализа и прогнозирования.

В Deductor Studio для аналитика основополагающим понятием является сценарий. Сценарий представляет собой последовательность операций с данными, представленную в виде иерархического дерева. В дереве каждая операция образует узел, заголовок которого содержит: имя источника данных, наименование применяемого метода обработки, используемые при этом поля и т.д. Кроме этого, слева от наименования узла стоит значок, соответствующий типу операции.

Если узел имеет подчиненные узлы, то слева от его названия будет расположен значок «+», щелчок по которому позволит развернуть узел, т.е. сделать видимыми все его подчиненные узлы, при этом значок «+» поменяется на «–». Щелчок по значку «–», наоборот, сворачивает все подчиненные узлы.

Задание

Ознакомиться с возможностями аналитического пакета Deductor, выполнив приведенные ниже задания. В конце работы сохранить проект.

Импорт данных

Сценарий состоит из ветвей. Deductor не имеет собственных средств для ввода данных, поэтому сценарий всегда начинается с узла импорта из какого-либо источника. Любой вновь создаваемый узел импорта будет находиться на верхнем уровне

Импорт данных из текстового файла с разделителями осуществляется путем вызова мастера импорта на панели «*Сценарии*» (рис. 4.1).

Рис. 4.1

После запуска мастера импорта указать тип импорта “*Текстовый файл (Direct)*” и перейти к настройке импорта. Все файлы примеров необходимо сохранить с сайта или взять с сетевого диска T:\Analiz.

Указать имя файла *Trade.txt* (), из которого необходимо получить данные. В данном файле содержатся данные о продажах за некоторый период времени. В окне просмотра выбранного файла можно увидеть его содержание.

Перейти к настройке параметров импорта (шаг 2–4). На шаге 3 указать в качестве разделителя целой и дробной части числа точку. Все остальные параметры по умолчанию оставить без изменения.

Теперь перейдем к настройке свойств полей. На этом шаге мастера предоставляется возможность настроить имя, название (**метку**), размер, тип данных, вид данных и назначение. Некоторые свойства (например, тип данных) можно задавать для выделенного набора столбцов. Вид данных определяет – конечный ли это набор (дискретные) или бесконечный (непрерывные). В нашем случае год и месяц продажи товара – это текстовые (строковые) дискретные значения, а количество проданного товара вещественные непрерывные.

Указав параметры столбцов, запустить процесс импорта, нажав на кнопку «Пуск».

После импорта данных на следующем шаге Мастера необходимо выбрать способ отображения данных. В данном случае самым информативным является диаграмма, поэтому выберем ее (рис. 4.2).

Рис. 4.2

Очистка данных

Часто исходные данные для анализа не годятся, а качество данных влияет на качество результатов. Обычно «сырые» данные содержат в себе различные «шумы», за которыми трудно увидеть общую картину, а также аномалии – влияние случайных, либо редко происходивших событий. Очевидно, что влияние этих факторов на общую модель необходимо минимизировать, т.к. модель, учитывающая их, получится неадекватной. Минимизировать влияние шумов, аномалий и прочее можно, используя устойчивые к их воздействию алгоритмы анализа и применяя специализированные механизмы очистки.

Парциальная предобработка

Парциальная предобработка служит для восстановления пропущенных данных, редактирования аномальных значений (выбросов) и спектральной

обработке данных (например, сглаживания данных). Именно этот шаг часто проводится в первую очередь.

Как видно из диаграммы (рис. 4.2) выбросы ухудшают статистическую картину распределения. Воспользуемся мастером обработки (рис. 4.3) и выберем *Редактирование выбросов*.

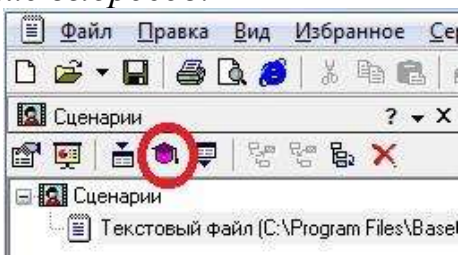


Рис. 4.3

В мастере на 2 шаге поставить флажок «Обрабатывать как упорядоченный набор данных». На 3 шаге для поля «*Количество*» указать большую степень подавления.

После выполнения процесса обработки на диаграмме (рис. 4.4) видно, что выбросы уменьшились, стала проясняться реальная картина продаж. Переключение между двумя диаграммами можно сделать через меню «*Окно*».

Рис. 4.4

Спектральная обработка

Сглаживание данных применяется для удаления шумов из исходного набора, а также для выделения тенденции, трудно обнаруживаемой в исходном наборе. Deductor Studio предлагает несколько видов спектральной обработки.

Продолжим работу с данными файла «*Trade.txt*». Сгладим данные при помощи спектральной обработки. В Мастере поле «*Количество*» указать

как используемое и выбрать «Вычитание шума» После выполнения процесса обработки выбрать в качестве визуализации диаграмму.

Как видно из рис. 4.5 данные стали более сглаженными и могут служить для дальнейшей обработки. Взглянув на данные легко понять общую тенденцию.

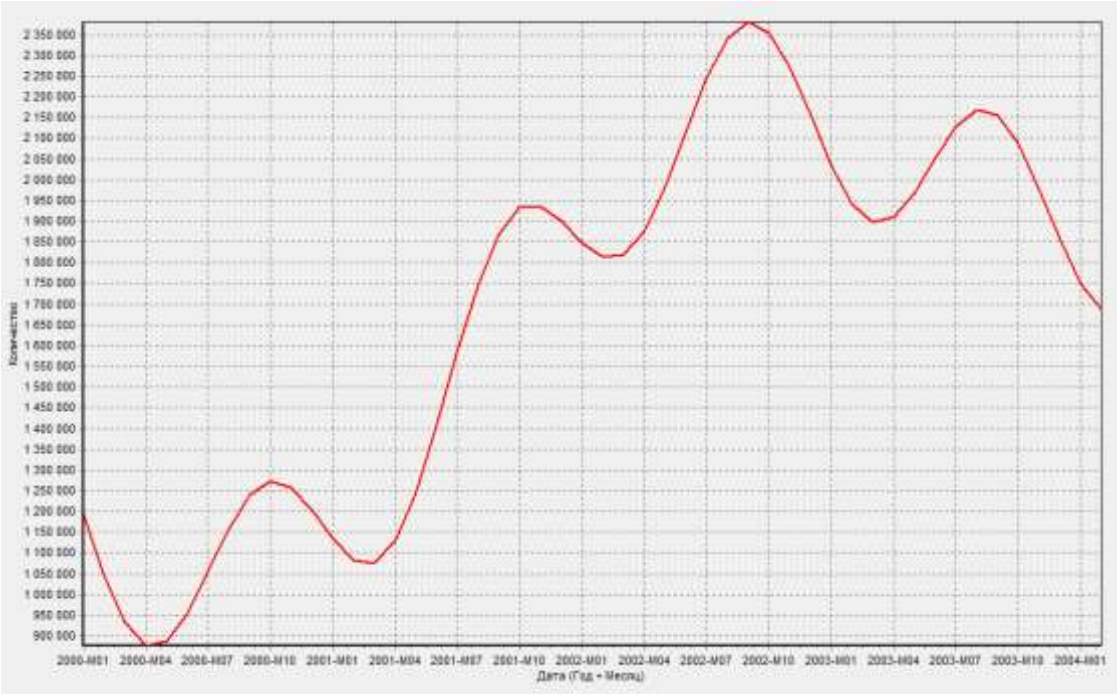


Рис. 4.5

Задание для самостоятельного выполнения

1. Импортировать сценарий «Dynamics_website.txt», содержащий данные о посещаемости веб-сайта.
2. Визуализировать данные в виде диаграммы.
3. Провести оценку качества данных по полю Посещения, используя соответствующий метод обработки.
4. Провести спектральную обработку по полю Посещения на основе Вейвлет-преобразования.
5. Визуализировать данные в виде диаграммы.
6. Провести сравнительный анализ диаграмм до и после обработки.

ЛАБОРАТОРНАЯ РАБОТА №5. Корреляционный анализ. Выявление дубликатов и противоречий данных

Задание

Ознакомиться с возможностями аналитического пакета Deductor, выполнив приведенные ниже задания. В конце работы сохранить проект.

Корреляционный анализ

Корреляционный анализ применяется для оценки зависимости выходных полей данных от входных факторов и устранения незначащих факто-

ров. Принцип корреляционного анализа состоит в поиске таких значений, которые в наименьшей степени коррелированны (взаимосвязаны) с выходным результатом. Такие факторы могут быть исключены из результирующего набора данных практически без потери полезной информации. Критерием принятия решения об исключении является порог значимости. Если корреляция между входным и выходным факторами меньше порога значимости, то соответствующий фактор отбрасывается как незначащий.

Рассмотрим применение обработки на примере данных из файла «*Anketa1.txt*». Он содержит таблицу с информацией о кредитах граждан. В данном примере определим степень влияния входных факторов на один из выходов и оставим только значимые.

Импортировать файл «*Anketa1.txt*» и выбрать вариант отображения таблица. В Мастере обработки выбрать корреляционный анализ, и задать входные поля «*Личный доход в месяц после налогообложения*», «*Сумма кредита*», «*Стаж работы*», «*Количество лет проживания в регионе*», «*Рыночная стоимость автомобиля*», «*Рыночная стоимость недвижимости*» и выходное поле «*Возврат кредита*». Остальные поля отметить как неиспользуемые (рис. 5.1).

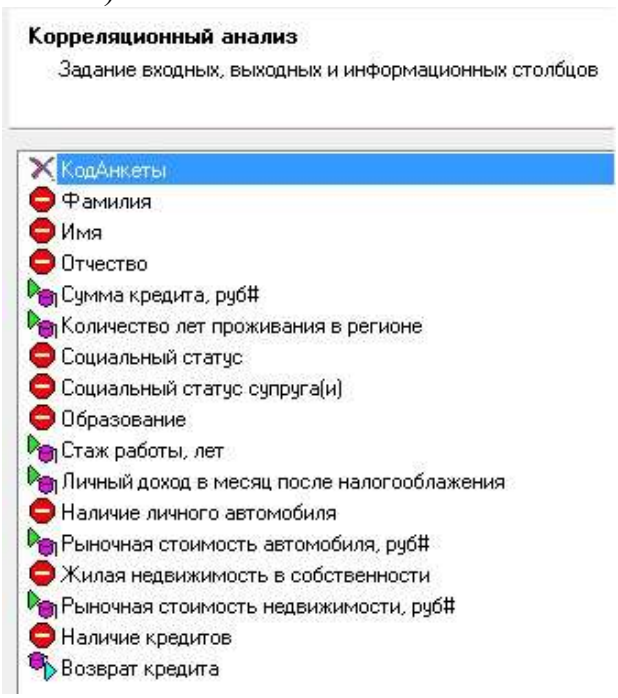


Рис. 5.1

В качестве метода выбрать коэффициент корреляции Пирсона.

На следующем шаге запустить процесс корреляционного анализа. После завершения процесса выбрать все факторы.

По полученной матрице корреляции (рис. 5.2) видно, какие факторы влияют сильнее, чем другие, и какие можно не учитывать при построении всевозможных моделей.

Входные поля		Корреляция с выходными полями
№	Поле	Возврат кредита
1	Сумма кредита, руб#	0,193
2	Количество лет проживания в регионе	-0,094
3	Стаж работы, лет	-0,010
4	Личный доход в месяц после налогообложения	-0,016
5	Рыночная стоимость автомобиля, руб#	-0,036
6	Рыночная стоимость недвижимости, руб#	-0,213

Рис. 5.2

Дубликаты и противоречия

Одна из серьезных проблем, часто встречающаяся на практике, – наличие в данных дубликатов и противоречий.

Противоречивыми являются группы записей, в которых содержатся строки с одинаковыми входными факторами, но разными выходными. В такой ситуации непонятно, какое результирующее значение верное. Если противоречивые данные использовать для построения модели, то она окажется неадекватной. Поэтому противоречивые данные чаще всего лучше вообще исключить из исходной выборки.

Также в данных могут встречаться записи с одинаковыми входными факторами и одинаковыми выходными, т.е. дубликаты. Таким образом, данные несут избыточность. В большинстве случаев дубликаты в данных являются следствием ошибок при подготовке данных.

В Deductor Studio для автоматизации этого процесса есть соответствующий инструмент – обработка «*Дубликаты и противоречия*».

Суть обработки состоит в том, что определяются входные (факторы) и выходные (результаты) поля. Алгоритм ищет во всем наборе записи, для которых одинаковым входным полям соответствуют одинаковые (дубликаты) или разные (противоречия) выходные поля. На основании этой информации создаются два дополнительных логических поля – «*Дубликат*» и «*Противоречие*», принимающие значения «истина» или «ложь». В дополнительные числовые поля "*Группа дубликатов*" и "*Группа противоречий*" записываются номер группы дубликатов и группы противоречий, в которые попадает данная запись. Если запись не является дубликатом или противоречием, то соответствующее поле будет пустым.

Рассмотрим механизм выявления дубликатов на примере данных файла «*Anketa.txt*». В этом файле находится информация об анкетных данных граждан, участвующих в кредитовании. Попробуем вычислить присутствие дубликатов.

Импортируем данные из текстового файла и посмотрим их в виде таблицы. Для выявления дубликатов запустить Мастер обработки. В нем выбрать тип обработки «*Дубликаты и противоречия*». На 2 шаге Мастера необходимо настроить назначение полей. Поля «*Фамилия*», «*Имя*», «*Отчество*» определить как входные, «*Код Анкеты*» – как выходное, а «*Сумма кредита*» оставить информационным.

После завершения выявления дубликатов просмотреть результат в виде таблицы дубликатов и противоречий.

В первом случае видно, что существуют одинаковые строки, являющиеся дубликатами. Данный обработчик показывает дубликаты и их принадлежность к группам дубликатов (рис. 5.3).

Дубликаты		Противоречия		Входные поля			Выходные поля	Информационные поля
Признак	Группа	Признак	Группа	Фамилия	Имя	Отчество	КодАнкеты	Сумма кредита, руб#
☑	1	☐		Бобров	Андрей	Владимирович	4076	105000
☑	2	☐		Калугин	Анатолий	Алексеевич	3056	58000
☑	3	☐		Полякова	Тамара	Ивановна	3076	36000
☑	4	☐		Широкова	Светлана	Николаевна	4000	54000
☑	1	☐		Бобров	Андрей	Владимирович	4076	105000
☑	4	☐		Широкова	Светлана	Николаевна	4000	54000
☑	3	☐		Полякова	Тамара	Ивановна	3076	36000
☑	2	☐		Калугин	Анатолий	Алексеевич	3056	58000

Рис. 5.3

Во втором случае видно, что при одинаковых «*Фамилия*», «*Имя*», «*Отчество*» оказываются различные «*Коды Анкет*» (рис. 5.4).

Дубликаты		Противоречия		Входные поля			Выходные поля	Информационные поля
Признак	Группа	Признак	Группа	Фамилия	Имя	Отчество	КодАнкеты	Сумма кредита, руб#
☐		☑	1	Абаев	Александр	Викторович	3061	32000
☐		☑	2	Евстафьев	Олег	Николаевич	4054	64000
☐		☑	1	Абаев	Александр	Викторович	4026	43000
☐		☑	2	Евстафьев	Олег	Николаевич	4039	47000
☐		☑	3	Ханнаков	Медахат	Рифкатович	4035	51000
☐		☑	3	Ханнаков	Медахат	Рифкатович	4013	120000

Рис. 5.4

Задание для самостоятельного выполнения

1. Импортировать сценарий «Region.txt», содержащий данные об экономическом состоянии регионов.
2. Провести корреляционный анализ зависимости объема Валового регионального продукта от всех остальных пригодных факторов.
3. Визуализировать данные в виде матрицы корреляции.
4. Сделать выводы по результатам анализа.

ЛАБОРАТОРНАЯ РАБОТА №6. Классификация с помощью деревьев решений

Задание

Ознакомиться с возможностями аналитического пакета Deductor, выполнив приведенные ниже задания. В конце работы сохранить проект.

Деревья решений.

Деревья решений применяются для решения задачи классификации. Дерево представляет собой иерархический набор условий (правил), согласно которым данные относятся к тому или иному классу. В построенном де-

реве присутствует информация о достоверности того или иного правила. Рассчитывается значимость каждого входного поля.

Пусть аналитик имеет данные по тому, как голосуют депутаты конгресса США по различным законопроектам. Также известна партийная принадлежность каждого депутата – республиканец или демократ. Перед аналитиком поставлена задача: классифицировать депутатов на демократов и республиканцев в зависимости от того, как они голосуют.

Данные по голосованию находятся в файле «*Голосование конгресса.txt*». Таблица содержит следующие поле «*Класс*» – класс голосующего (демократ или республиканец) и поля, информирующие о том, как голосовали депутаты за принятие различных законопроектов («да», «нет», «воздержался»).

Для решения задачи нужно импортировать файл *Голосование конгресса.txt* (все типы полей указать как строковые), и запустить Мастер обработки. Выбрать в качестве обработки дерево решений. В Мастере построения на 2 шаге сделать поле «*Класс*» выходным, а остальные поля входными. Далее предлагается настроить способ разбиения исходного множества данных на обучающее и тестовое. Зададим случайный способ разбиения, когда данные для тестового и обучающего множества берутся из исходного набора случайным образом.

На следующем шаге Мастера предлагается настроить параметры процесса обучения, а именно минимальное количество примеров, при котором будет создан новый узел (пусть узел создается, если в него попали два и более примеров), а также предлагается возможность строить дерево с более достоверными правилами.

На следующем шаге Мастера запускается сам процесс построения дерева. Также можно увидеть информацию о количестве распознанных примеров (рис. 6.1).

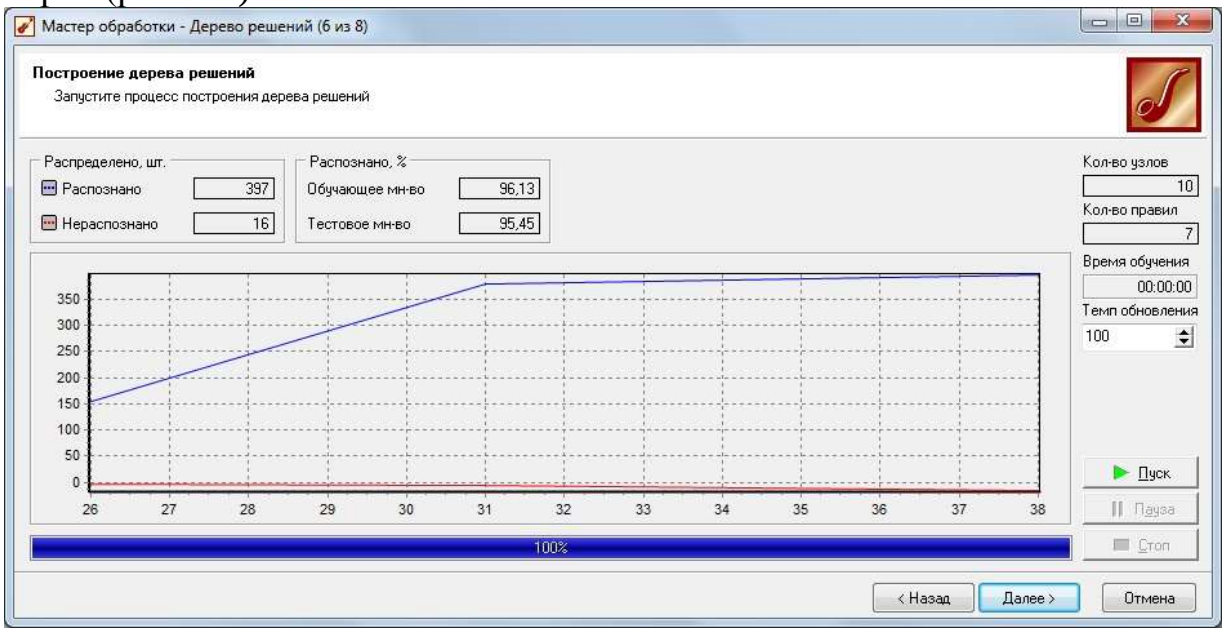


Рис. 6.1

После построения дерева можно увидеть, что почти все примеры и на обучающей и на тестовой выборке распознаны.

Перейти на следующий шаг Мастера для выбора способа визуализации. Основной целью аналитика является отнесение депутата к той или иной партии. Механизм отнесения должен быть таким, чтобы депутат указал, как он будет голосовать за различные законопроекты, а дерево решений ответит на вопрос, кто он – демократ или республиканец. Такой механизм предлагает визуализатор «*Что-если*».

Не менее важным является и просмотр самого дерева решений, на котором можно определить, какие факторы являются более важными (верхние узлы дерева), какие второстепенными, а какие вообще не оказывают влияния (входные факторы, вообще не присутствующие в дереве решений). Поэтому выберем также и визуализатор «*Дерево решений*».

Формализованные правила классификации, выраженные в форме «Если <Условие>, тогда <Класс>», можно увидеть, выбрав визуализатор «*Правила (дерево решений)*».

Часто аналитику бывает полезно узнать, сколько примеров было распознано неверно, какие именно примеры были отнесены к какому классу ошибочно. На этот вопрос дает ответ визуализатор «*Таблица сопряженности*».

Важную информацию предоставляет визуализатор «*Значимость атрибутов*». С помощью него можно определить, насколько сильно выходное поле зависит от каждого из входных факторов. Чем больше значимость атрибута, тем больший вклад он вносит при классификации.

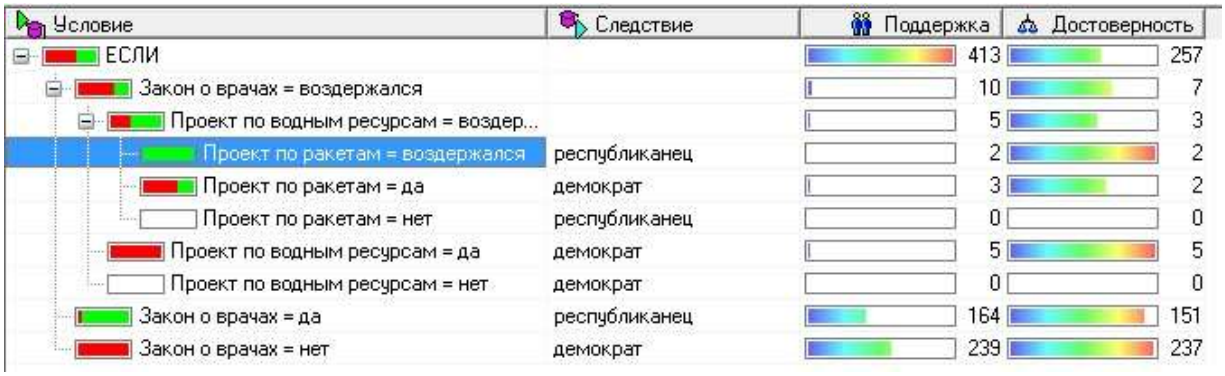
Проведем анализ полученных данных. Для начала посмотрим на «*Таблицу сопряженности»* (рис. 6.2).

	Классифицировано		
Фактически	демократ	республиканец	Итого
демократ	253	14	267
республиканец	3	165	168
Итого	256	179	435

Рис. 6.2

По диагонали таблицы расположены примеры, которые были правильно распознаны, в остальных ячейках – те, которые были отнесены к другому классу. В данном случае дерево правильно классифицировало практически все примеры.

Перейдем к визуализатору «*Дерево решений»* (рис. 7.3). Как видно, дерево решений получилось не очень громоздкое, большая часть факторов (законопроектов) была отсечена, т.е. влияние их на принадлежность к партии минимальна или его вообще нет (по-видимому, по этим вопросам у партий нет принципиального противостояния).



Условие	Следствие	Поддержка	Достоверность
ЕСЛИ		413	257
Закон о врачах = воздержался		10	7
Проект по водным ресурсам = воздер...		5	3
Проект по ракетам = воздержался	республиканец	2	2
Проект по ракетам = да	демократ	3	2
Проект по ракетам = нет	республиканец	0	0
Проект по водным ресурсам = да	демократ	5	5
Проект по водным ресурсам = нет	демократ	0	0
Закон о врачах = да	республиканец	164	151
Закон о врачах = нет	демократ	239	237

Рис. 6.3

Самым значимым фактором оказалась позиция, занимаемая депутатами по пакету законов, касающихся врачей. Это же подтверждает и визуализатор «*Значимость атрибутов*».

На визуализаторе «*Правила*» представлен список всех правил, согласно которым можно отнести депутата к той или иной партии. Правила можно сортировать по поддержке, достоверности, фильтровать по выходному классу (к примеру, показать только те правила, согласно которым депутат является демократом с сортировкой по поддержке).

Данные представлены в виде таблицы. Полями этой таблицы являются:

- номер правила,
- условие, которое однозначно определяет принадлежность к партии,
- решение – то, кем является депутат, голосовавший согласно этому условию,
- поддержка – количество и процент примеров из исходной выборки, которые отвечают этому условию,
- достоверность – процентное отношение количества верно распознанных примеров, отвечающих данному условию, к общему количеству примеров, отвечающих данному условию.

Исходя из данных этой таблицы, аналитик может сказать, что именно влияет на то, что депутат является демократом или республиканцем, какова цена этого влияния (поддержка) и какова достоверность правила.

Задание для самостоятельного выполнения

Вариант 1 (номер зачетной книжки заканчивается на 0, 1, 2, 3)

На основе исходных данных файла «Credit.txt» определить правила выдачи кредитов.

Вариант 2 (номер зачетной книжки заканчивается на 4, 5, 6)

На основе исходных данных файла «Грибы.txt» найти алгоритм определения съедобных и не ядовитых грибов.

Вариант 3 (номер зачетной книжки заканчивается на 7, 8, 9)

На основе исходных данных файла «Ирисы.txt» найти алгоритм определения класса цветов.

ЛАБОРАТОРНАЯ РАБОТА №7. Прогнозирование с помощью линейной регрессии

Задание

Ознакомиться с возможностями аналитического пакета Deductor, выполнив приведенные ниже задания. В конце работы сохранить проект.

Определение сезонности

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

Рассмотрим применение линейной регрессии на примере данных по продажам, находящихся в файле «Trade.txt».

Выполните импорт данных из файла «*Trade.txt*», не забыв указать в Мастере, чтобы в качестве разделителя дробной и целой части была точка, а не запятая.

Одна из основных целей анализа временных рядов – это возможность построения прогноза. Наиболее простой метод прогнозирования значения ряда – это экстраполяция. Однако она возможна лишь при достаточной закономерности развития изучаемого явления. Реальные экономические процессы редко в достаточной степени удовлетворяют этому требованию. Поэтому исходный временной ряд представляют как совокупность нескольких компонент, к каждой из которых применяется свой метод прогнозирования в соответствии с тенденциями, установленными в прошлом.

Целью декомпозиции временного ряда является выделение и изучение сезонной составляющей и тренда. Построим тренд и проверим наличие сезонности продаж. Для этого воспользуемся обработчиком «*Декомпозиция временного ряда*». Как видно из диаграммы декомпозиции (рис. 7.1), присутствует сезонность с периодом 12 месяцев.

Преобразование данных к скользящему окну

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

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

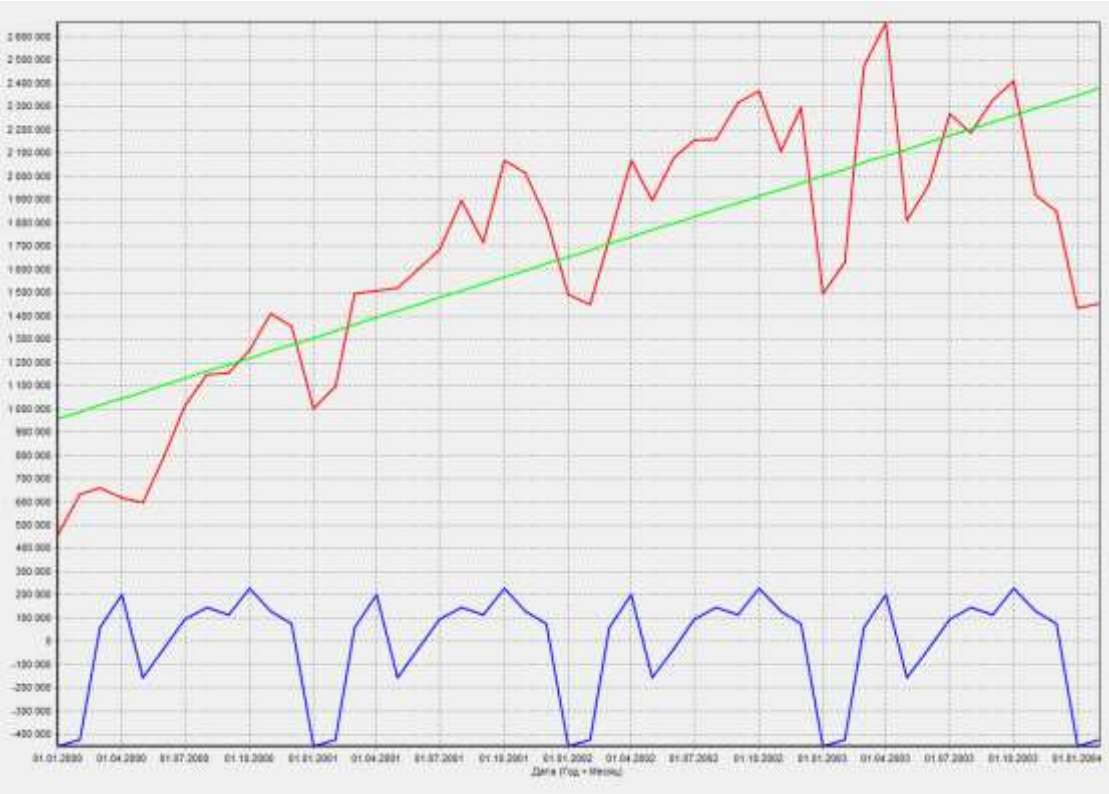


Рис. 7.1

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.

Запустите Мастер обработки «*Скользящее окно*» (рис. 7.2).

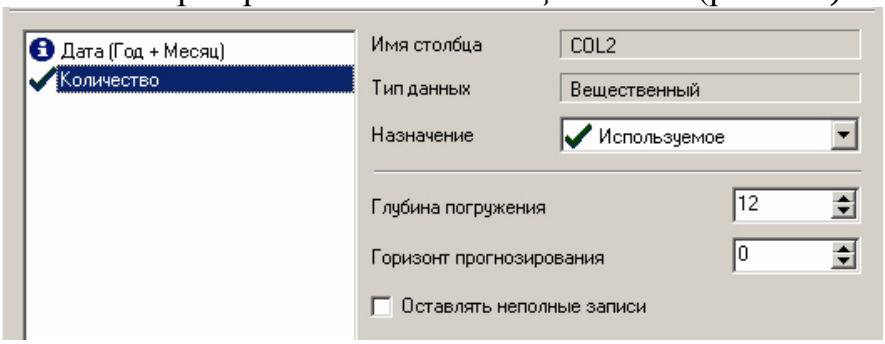


Рис. 7.2

Как видно, теперь в качестве входных факторов можно использовать "Количество - 12", "Количество - 11" – данные по количеству 12 и 11 месяцев назад (относительно прогнозируемого месяца) и остальные необходимые факторы. В качестве результата прогноза будет указан столбец "Количество" (рис. 7.3).

	Дата (Год + Месяц)	Количество-12	Количество-11	Количество-10	Количество-9	Количество-8
▶	2001-M02	633208,196	660159,299	617455,3417	597354,4794	793517,4512
	2001-M03	660159,299	617455,3417	597354,4794	793517,4512	1015944,2862
	2001-M04	617455,3417	597354,4794	793517,4512	1015944,2862	1148052,2523
	2001-M05	597354,4794	793517,4512	1015944,2862	1148052,2523	1156623,1715
	2001-M06	793517,4512	1015944,2862	1148052,2523	1156623,1715	1255021,9423
	2001-M07	1015944,2862	1148052,2523	1156623,1715	1255021,9423	1410114,5606
	2001-M08	1148052,2523	1156623,1715	1255021,9423	1410114,5606	1357230,3388
	2001-M09	1156623,1715	1255021,9423	1410114,5606	1357230,3388	1003317,7317
	2001-M10	1255021,9423	1410114,5606	1357230,3388	1003317,7317	1097048,6263
	2001-M11	1410114,5606	1357230,3388	1003317,7317	1097048,6263	1498977,3427
	2001-M12	1357230,3388	1003317,7317	1097048,6263	1498977,3427	1507696,4482

Рис. 7.3

Прогнозирование с помощью линейной регрессии

Теперь можно перейти непосредственно к прогнозированию. Будем строить прогноз с помощью линейной регрессии сразу после обработчика «*Скользящее окно*».

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

На первом шаге задать назначение исходных столбцов. «*Количество 1–12»* сделать входными столбцами, а «*Год и месяц»* информационным полем. В качестве выходного поля указать столбец «*Количество*».

На следующем шаге происходит настройка обучающего и тестового множеств, способ разложения исходного множества данных. Третий шаг установки позволяет осуществить ограничение диапазона входных значений. Данные шаги оставим без изменений. При нажатии на кнопку «*Далее*» появляется окно запуска процесса обучения. В процессе выполнения видно, какая часть распознана на этапе обучения и теста.

После выполнения процесса выбрать в качестве способа отображения диаграмму рассеяния и отображение результатов в виде диаграммы. Как видно из обеих диаграмм, обучение прошло с хорошей точностью (рис. 7.4).

Теперь для построения прогноза запустить Мастера обработки, в котором выбрать *Прогнозирование*. На первом шаге обработчика происходит настройка связи столбцов для прогнозирования. Указать связь между столбцами и горизонт прогноза равный 3 (рис. 7.5).

На следующем шаге задаются параметры визуализации. Для данного примера выбрать отображение результатов в виде диаграммы прогноза (рис. 7.6).

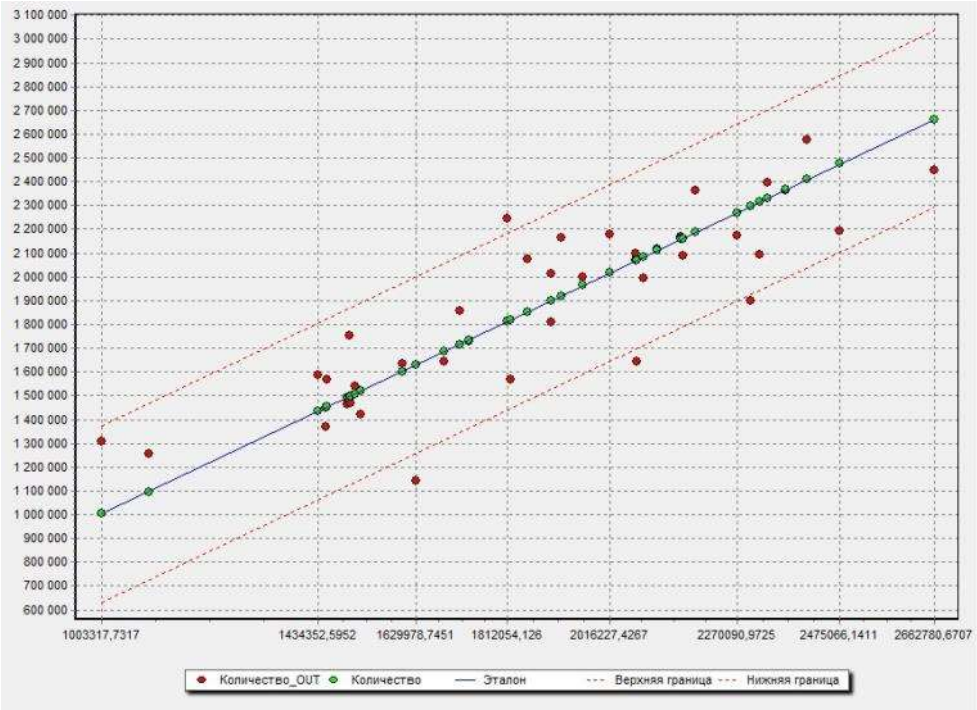


Рис. 7.4

Теперь аналитик может дать прогноз о продажах, основываясь на модели, построенной с помощью линейной регрессией.

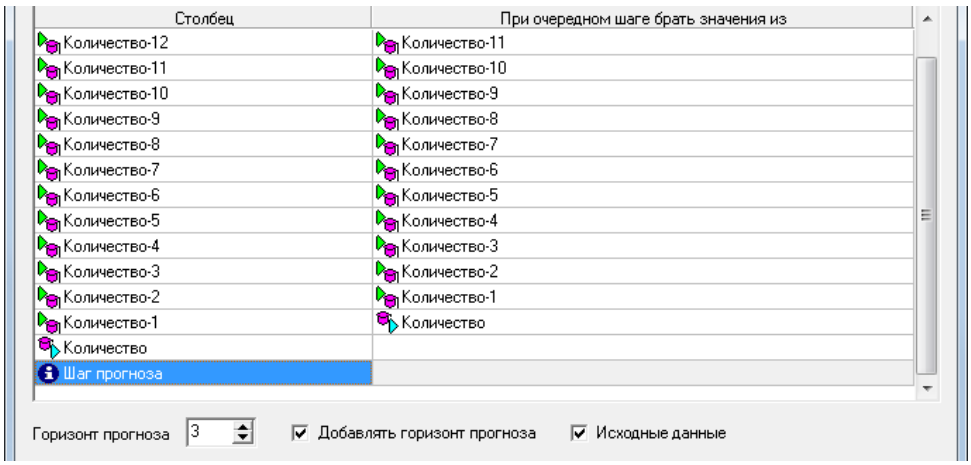


Рис. 7.5

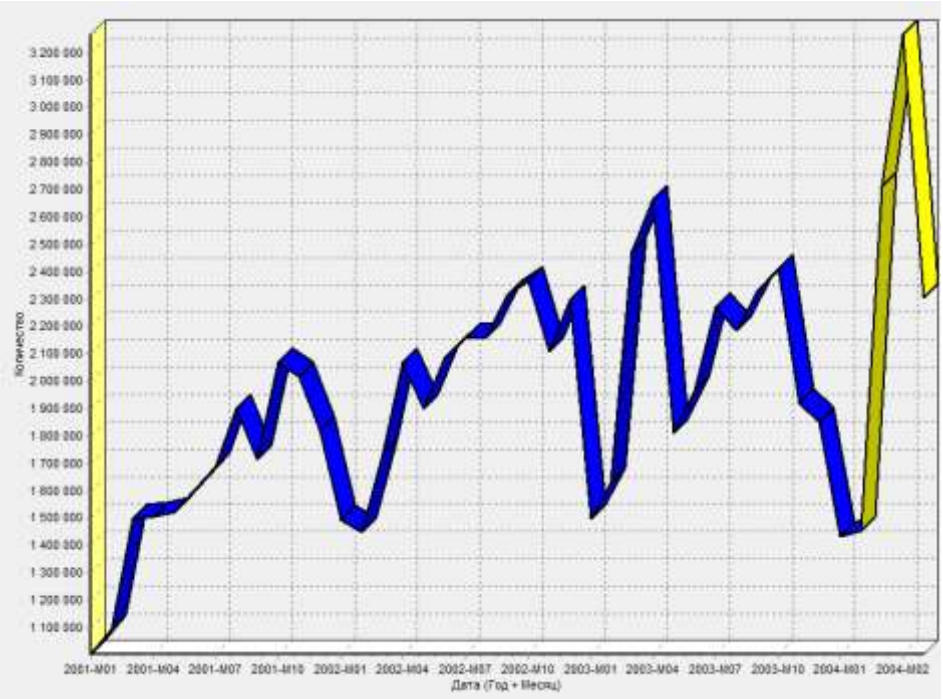


Рис. 7.6

Задание для самостоятельного выполнения

1. Используя результаты файл «dynamics_website.txt» построить скользящее окно с глубиной погружения 12.
2. Построить и оценить диаграмму рассеяния.
3. Осуществить прогнозирование на 6 периодов.
4. Визуализировать данные в виде гистограммы.
5. Сделать выводы по результатам анализа.

ЛАБОРАТОРНАЯ РАБОТА №8. Кластеризация данных

Задание

Ознакомиться с возможностями аналитического пакета Deductor, выполнив приведенные ниже задания. В конце работы сохранить проект.

Кластеризация с помощью с помощью алгоритма k-means

Рассмотрим механизм кластеризации, реализованный на алгоритме k-means, основываясь на данных роста численности населения по регионам РФ. Исходная таблица находится в файле «*Регионы.txt*». Задача состоит в распределении регионов на функциональные группы по демографической картине в них и выявлении скрытых закономерностей.

Вначале необходимо осуществить импорт рассматриваемых данных из файла. После этого запустить Мастер обработки «*Кластеризация*». При запуске Мастера необходимо настроить назначения столбцов, т.е. выбрать свойства, по которым будет происходить группировка объектов. Укажите

столбцам «*Численность населения*» и «*Регион*» назначение «*информационное*», а остальным полям – «*входное*».

На следующем шаге Мастера необходимо настроить способ разделения исходного множества данных на тестовое и обучающее, а также количество примеров в том и другом множестве. Укажите, что данные обоих множеств берутся случайным образом, и определите все множество как обучающее.

Следующий шаг предлагает настроить параметры кластеризации, определить на какое количество кластеров будет распределяться исходное множество. По мнению экспертов в стране наблюдается четыре тенденции развития регионов, поэтому выберем фиксированное количество кластеров равное четырем.

Для отображения полученных групп кластеров выбрать из списка визуализаторов способы отображения данных: «*Профили кластеров*», «*Куб*» «*Матрица сравнений», «Связи кластеров».*

Для настройки визуализатора «*Куб»* необходимо выбрать рассматриваемые показатели как факты, а номер кластера и регионы как измерение.

На 9 шаге задать отображение фактов как среднее по рассматриваемое группе.

Общую структуру сформированных алгоритмом кластеров можно просмотреть в визуализаторе «*Профили кластеров*» (рис. 8.1). В нем представлены все рассматриваемые свойства вместе с характером влияния их на состав кластера. Основным фактором, определяющим состав кластера, является значимость свойств, выраженная в процентах.

Алгоритм автоматически разбил регионы на четыре кластера с разной поддержкой и разными процентами значимости свойств.

		Кластеры				
		2	1	3	0	Итого
		47 (52,8%)	32 (36,0%)	8 (9,0%)	2 (2,2%)	
Поля	Показатели					
9.0 Среднедуш	Значимость	100,0%	78,9%	100,0%	100,0%	100,0%
9.0 Среднегод	Значимость	41,9%	46,5%	74,6%	73,1%	68,7%
9.0 Число	Значимость	53,2%	25,8%	67,1%	70,1%	63,1%
9.0 Удельный	Значимость	23,9%	27,7%	71,3%	77,5%	62,4%
9.0 Число	Значимость	38,6%	55,4%	7,0%	23,8%	24,3%
9.0 Изменение	Значимость	23,3%	37,0%	39,1%	2,7%	9,3%

Рис. 8.1

Второй кластер является показателем демографической обстановки страны, так как собрал в себя максимальное количество регионов – 47 из 89.

Малозначимым и почти не влияющим свойством на распределение является изменение численности населения по сравнению с предыдущим годом, при необходимости данным свойством можно пренебречь.

С помощью кнопки переименование кластеров можно им присвоить им рабочие названия (рис. 8.2).

Переименование кластеров

№	Название
1	Вымирающие большие регионы
2	Дотационные возрождающиеся
3	Возрождающиеся
4	Вымирающие малые регионы

Рис. 8.2

Наиболее ярко выраженными кластерами по заданным свойствам являются нулевой и первый кластер: они максимально отличаются от остальных рассматриваемых групп значениями свойств, и минимальной поддержкой. Подтвердим предположение, используя визуализатор «*Матрица сравнений*». Наименьшая степень сходства между первым и нулевым кластером 12,76% (рис. 8.3).

	0	1	2	3
0	100,00%	12,76%	0,00%	36,05%
1	12,76%	100,00%	87,24%	76,71%
2	0,00%	87,24%	100,00%	63,95%
3	36,05%	76,71%	63,95%	100,00%

Рис. 8.3

Так же оценить похожесть кластеров можно с помощью визуализатора «*Связи кластеров*». Наиболее похожим на нулевой кластер является второй, он имеет наибольшую степень связи, отображаемую на диаграмме красным цветом (рис. 9.4). При необходимости данные кластеры как наиболее похожие можно объединить.

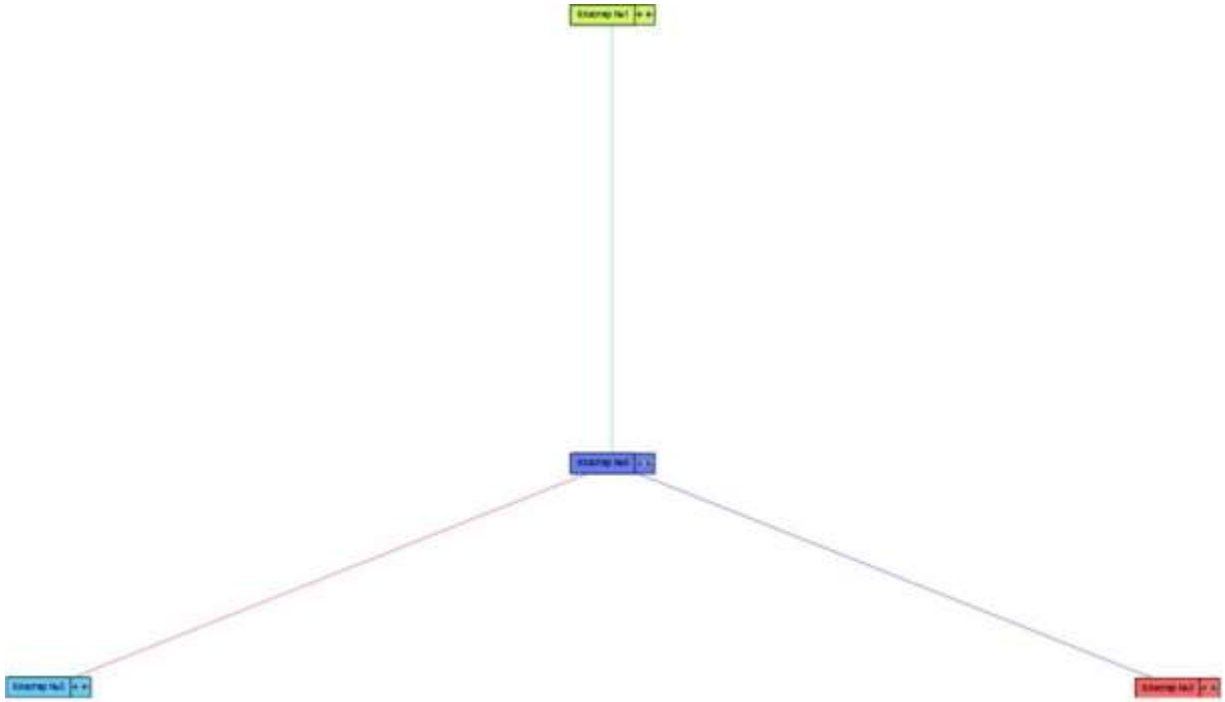


Рис. 8.4

Результаты по сформированным кластерам наиболее удобно рассматриваются с помощью визуализатора «*Куб*», в который встроена кросс-диаграмма, изображающая полученные кластеры в графическом виде, что существенно упрощает анализ (рис. 8.5).

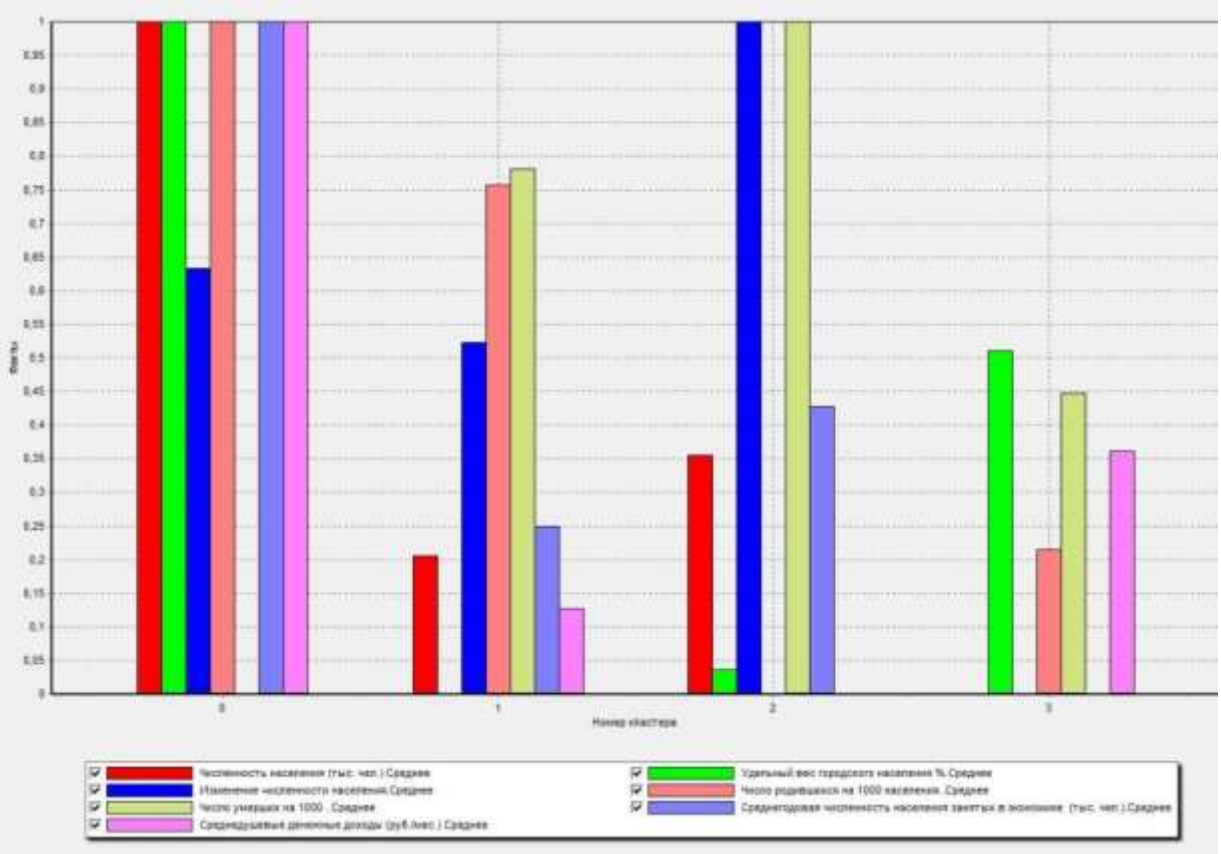


Рис. 8.5

Определить какой регион максимально похож на выбранный можно с помощью «*Диаграммы связи*». Определим семь похожих на Санкт-Петербург регионов по демографической обстановке. Зададим количество связей равным - 7, а Санкт-Петербург поместим в центр. Челябинская область максимально схожа с Санкт-Петербургом, степень сходства – 91,6%. В правой стороне окна можно проанализировать демографические коэффициенты для этих двух регионов (рис. 8.6).

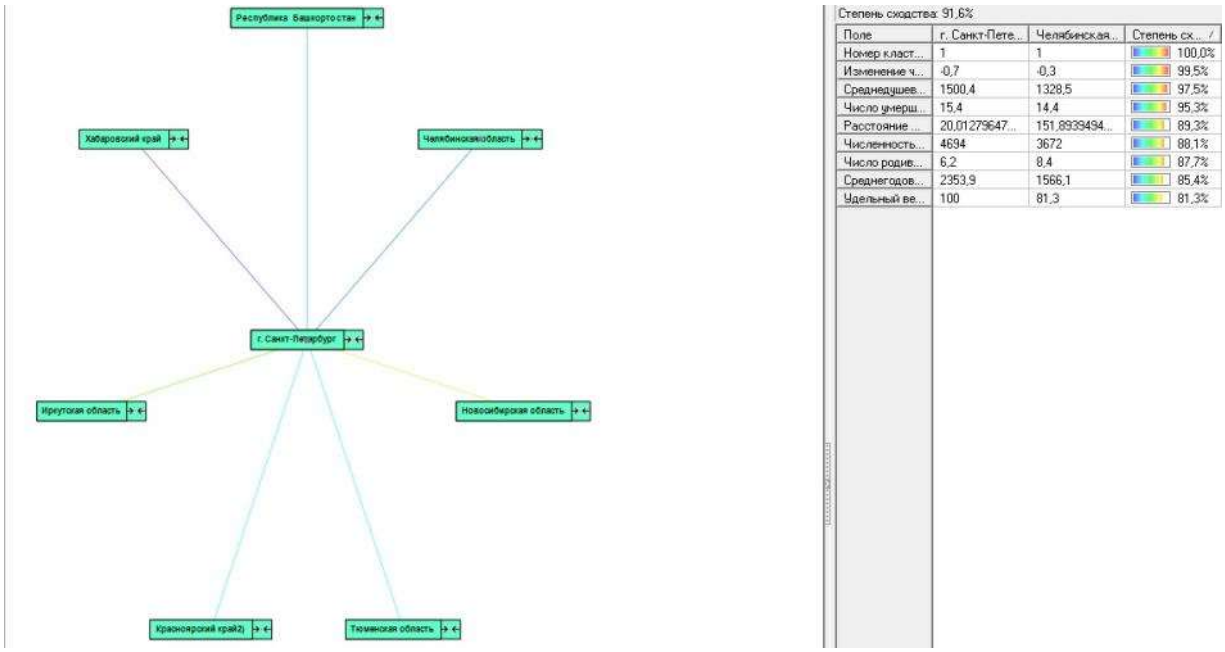


Рис. 8.6

Кластеризация с помощью самоорганизующейся карты Кохонена

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.

Рассмотрим механизм кластеризации путем построения самоорганизующейся карты, основываясь на тех же исходных данных о регионах.

Запустите Мастер обработки и выберите метод обработки «*Карта Кохонена*». Все поля кроме Региона (информационное поле) сделать входными.

На 3 шаге Мастера необходимо настроить способ разделения исходного множества данных на тестовое и обучающее, а также количество примеров в том и другом множестве. Укажите, что данные обоих множеств берутся случайным образом.

Следующий шаг предлагает настроить параметры карты. Значения по умолчанию вполне подходят. На 5 шаге Мастера также оставим параметры по умолчанию.

На 6 шаге настраиваются остальные параметры обучения. Укажем фиксированное количество кластеров – 4.

На 7 шаге предлагается запустить сам процесс обучения. Во время обучения можно посмотреть количество распознанных примеров и текущие значения ошибок. Нажмите кнопку «*Пуск*» и дождитесь завершения процесса обработки.

После этого требуется в списке визуализаторов выбрать появившуюся теперь *Карту Кохонена* для просмотра результатов кластеризации и *Профили кластеров*.

Далее в Мастере настройки отображения карты Кохонена указать все входные столбцы.

В итоге получаем Карту Кохонена (рис. 8.7), позволившую представить многомерное пространство входных факторов в двумерном виде, который удобнее анализировать.

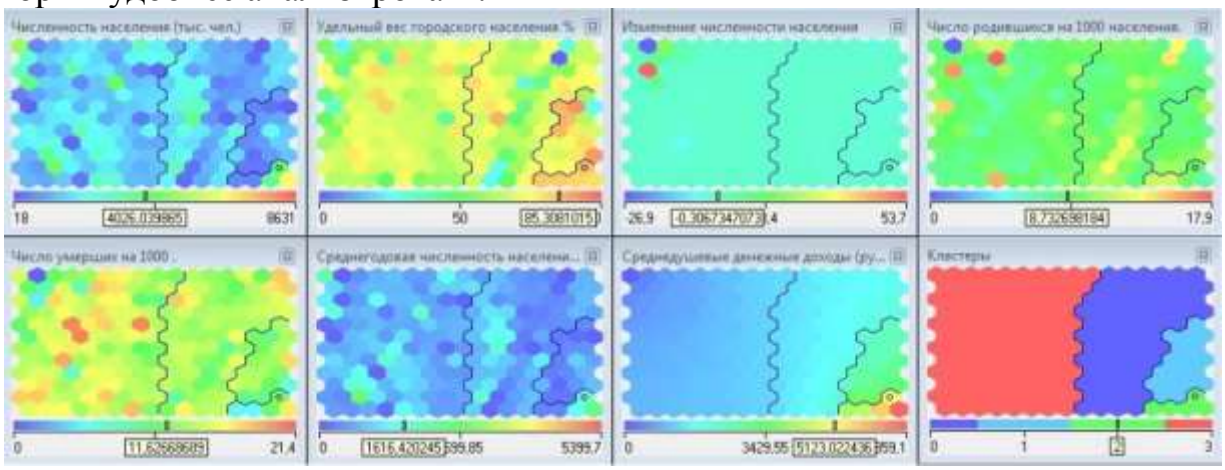


Рис. 8.7

Как и в случае кластеризации методом k-means, видно, что изменение численности населения не влияет на разбиение регионов. Наиболее эффективным кластером является кластер 2, объединяющий в себе возрождающиеся регионы с самыми высокими среднедушевыми доходами и относительно низкой смертностью.

Аналогичную информацию предоставляют нам и визуализатор «*Профили кластеров*».

Задание для самостоятельного выполнения

1. Импортировать сценарий «Абоненты.txt», с целью сегментирования абонентов телекоммуникационной компании, предоставляющей на рынке услуги мобильной связи.

2. Используя Карту Кохонена, сегментировать клиентов по семи кластерам.

2.1. На 2 шаге в «Настройках нормализации» для всех полей кроме Возраст и Среднемесячный расход (100%) установить значимость равную 30%.

2.2. Тестовое множество не использовать.

2.3. Увеличить размер карты в 1,5 раза. При размере 24x18 она будет иметь количество ячеек 432, в среднем на одну ячейку будет приходиться 10 примеров.

3. Визуализировать данные в виде Связей кластеров, Профилей кластеров и Карты Кохонена и проанализировать результат.

4. Переименовать кластеры задав им следующие названия: Бизнес-люди, Тусовщики, Работающие люди неактивные, Молодежь неактивная, Активная группа зрелого и пенсионного возраста, Группа предпенсионного возраста, неактивная, Группа пенсионного возраста, неактивная.

ЛАБОРАТОРНАЯ РАБОТА №9. Поиск ассоциативных правил

Поиск ассоциативных правил

Ассоциативные правила позволяют находить закономерности между связанными событиями. Примером такого правила служит утверждение, что покупатель, приобретающий хлеб, приобретет и молоко.

Пусть имеется база данных, состоящая из покупательских транзакций. Каждая транзакция – это набор товаров, купленных покупателем за один визит. Такую транзакцию еще называют рыночной корзиной. Целью анализа является установление следующих зависимостей: если в транзакции встретился некоторый набор элементов X, то на основании этого можно сделать вывод о том, что другой набор элементов Y также должен появиться в этой транзакции. Установление таких зависимостей дает нам возможность находить очень простые и интуитивно понятные правила.

Рассмотрим механизм поиска ассоциативных правил на примере данных о продажах товаров в некоторой торговой точке. Данные находятся в файле «*Supermarket.txt*». В таблице представлена информация по покупкам продуктов нескольких групп. Она имеет всего два поля «*Номер чека*» и «*Товар*». Необходимо решить задачу анализа потребительской корзины с целью последующего применения результатов для стимулирования продаж.

При импорте сценария указать, что поле «*Номер чека*» должно быть дискретным.

Для поиска ассоциативных правил запустить Мастер обработки и выбрать тип обработки «*Ассоциативные правила*». Далее указать, что поле «*Номер чека*» является идентификатором транзакции, а «*Товар*» элементом транзакции.

Следующий шаг позволяет настроить параметры построения ассоциативных правил: минимальную и максимальную поддержку, минимальную и максимальную достоверность, а также максимальную мощность множества. Исходя из характера имеющихся данных, следует указать границы поддержки – 13% и 80% и достоверности 60% и 90%.

После завершения процесса поиска, полученные результаты можно посмотреть, используя появившиеся специальные визуализаторы «*Популярные наборы*», «*Правила*», «*Дерево правил*», «*Что-если*».

Популярные наборы – это множества, состоящие из одного и более элементов, которые наиболее часто встречаются в транзакциях одновременно. Насколько часто встречается множество в исходном наборе транзакций, можно судить по поддержке. Данный визуализатор отображает множества в виде списка (рис. 9.1).

№	Множество	Поддержка	
		%	Кол-во
7	ЧАЙ	75,00	33
3	МАКАРОННЫЕ ИЗДЕЛИЯ	54,55	24
2	КЕТЧУПЫ, СОУСЫ, АДЖИКА	52,27	23
4	МЕД	50,00	22

Рис. 9.1

Получившиеся наборы товаров наиболее часто покупают в данной торговой точке, следовательно, можно принимать решения о поставках товаров, их размещении и т.д.

Визуализатор «*Правила*» отображает ассоциативные правила в виде списка правил (рис. 9.2). Таким образом, эксперту предоставляется набор правил, которые описывают поведение покупателей.

№	Условие	Следствие	Поддержка		Д
			%	Кол-во	
1	ВАФЛИ	СУХАРИ	22,73	10	
2	СУХАРИ	ВАФЛИ	22,73	10	
3	КЕТЧУПЫ, СОУСЫ, АДЖИКА	МАКАРОННЫЕ ИЗДЕЛИЯ	45,45	20	
4	МАКАРОННЫЕ ИЗДЕЛИЯ	КЕТЧУПЫ, СОУСЫ, АДЖИКА	45,45	20	
5	МЕД	ЧАЙ	40,91	18	

Рис. 9.2

Визуализатор «*Дерево правил*» – это всегда двухуровневое дерево. Оно может быть построено либо по условию, либо по следствию. При построении дерева правил по условию на первом (верхнем) уровне находятся узлы с условиями, а на втором уровне – узлы со следствием. Второй вариант дерева правил – дерево, построенное по следствию. Здесь на первом уровне располагаются узлы со следствием.

Справа от дерева находится список правил, построенный по выбранному узлу дерева. Для каждого правила отображаются поддержка и достоверность. Если дерево построено по условию, то вверху списка отображается условие правила, а список состоит из его следствий. Тогда правила отвечают на вопрос, что будет при таком условии. Если же дерево построено по следствию, то вверху списка отображается следствие правила, а список состоит из его условий. Эти правила отвечают на вопрос, что нужно, чтобы было заданное следствие.

В данном случае правила отображены по условию (рис. 9.3).

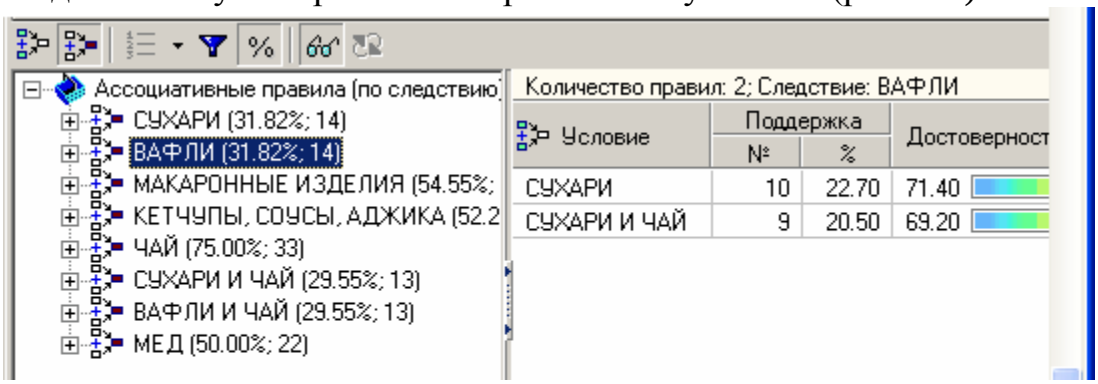


Рис. 9.3

Отображаемый результат можно интерпретировать как 2 правила:

1. Если покупатель приобрел вафли, то он с вероятностью 71% также приобретет сухари.

2. Если покупатель приобрел вафли, то он с вероятностью 64% также приобретет сухари и чай.

Анализ «*Что-если*» позволяет ответить на вопрос, что получим в качестве следствия, если выберем данные условия? Например, какие товары приобретаются совместно с выбранными товарами. В окне слева расположен список всех элементов транзакций. Справа от каждого элемента указана поддержка: сколько раз данный элемент встречается в транзакциях.

В правом верхнем углу расположен список элементов, входящих в условие. Это, например, список товаров, которые приобрел покупатель. Для них нужно найти следствие. Например, товары, приобретаемые совместно с ними. Чтобы предложить человеку то, что он, возможно, забыл купить.

В правом нижнем углу расположен список следствий. Справа от элементов списка отображается поддержка и достоверность.

Пусть необходимо проанализировать, что, возможно, забыл покупатель приобрести, если он уже взял вафли и мед. Для этого следует добавить в список условий эти товары (например, с помощью двойного щелчка мыши) и затем нажать на кнопку «*Вычислить правила*» (Ctrl+Enter). При этом в списке следствий появятся товары, совместно приобретаемые с данными. В данном случае появятся «сухари», «чай», «сухари и чай», т. е., может быть, покупатель забыл приобрести сухари, чай или и то и другое (рис. 9.4).

Таким образом, в данном примере найденные правила можно использовать для сегментации клиентов на два сегмента: клиенты, покупающие макаронные изделия и соусы к ним, и клиенты, покупающие все к чаю. В разрезе анализа предпочтений можно узнать, что наибольшей популярностью в данном магазине пользуются чай, мед, макаронные изделия, кетчупы, соусы и аджика. В разрезе размещения товаров в супермаркете можно применить результаты предыдущих двух анализов, т. е. располагать чай

рядом с медом, а кетчупы, соусы и аджику рядом с макаронными изделиями и т.д.

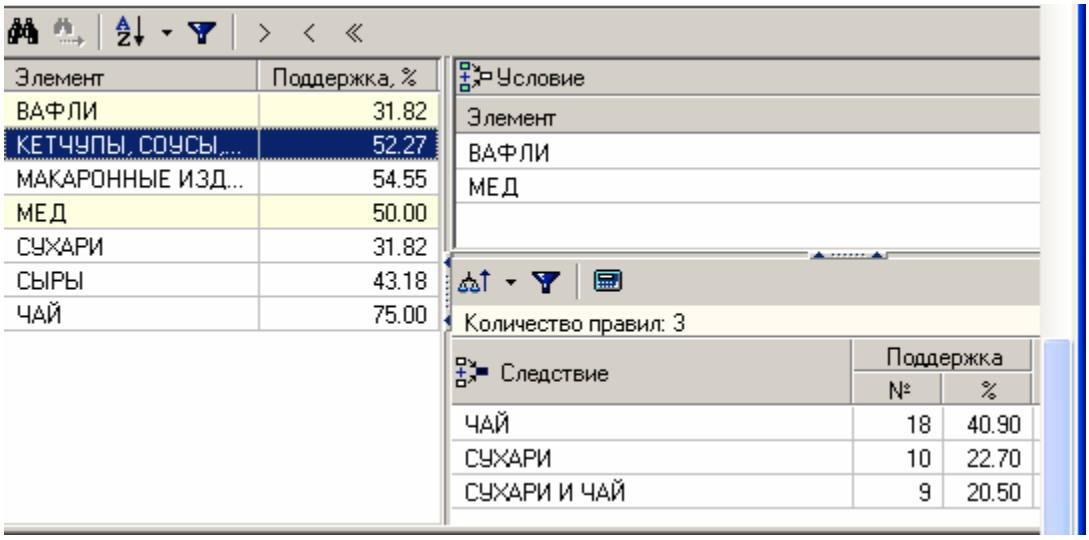


Рис. 9.4

Задание для самостоятельного выполнения

1. Импортировать сценарий «Чеки.txt», содержащий информацию из розничной сети продаж бытовой химии.

2. Проанализировать полученную информацию с помощью визуализаторов «Правила», «Дерево правил», «Что-если», «Популярные наборы».

ЛАБОРАТОРНАЯ РАБОТА №10. Предсказательная аналитика с помощью нейронной сети

Задание на выполнение работы.

Ознакомиться с возможностями аналитического пакета Deductor, выполнив приведенные ниже задания. В конце работы сохранить проект.

Прогнозирование с помощью нейронной сети

Особенностью процесса оценки стоимости объекта имущества является его рыночный характер. Это означает, что процесс оценки объекта не ограничивается учетом одних только затрат на создание или приобретение оцениваемого объекта собственности - необходим учет совокупности рыночных факторов, экономических особенностей оцениваемого объекта, а также макроэкономического и микроэкономического окружения. Кроме того, рынок недвижимости очень динамичный, поэтому требуется периодическая переоценка объектов собственности.

Нейросети как универсальные аппроксиматоры позволяют строить сложные нелинейные регрессионные модели типа "черный ящик". Создание моделей для оценки стоимости недвижимости могут существенно повысить эффективность работы организаций, занимающихся риэлтерской деятельностью.

Рассмотрим данный механизм на примере таблицы продаж из файла «*Недвижимость.txt*». При импорте обратите внимание на типы и виды числовых данных (при необходимости их нужно изменить).

Для построения модели использовались данные по стоимости квартир на вторичном рынке жилья одного из крупных городов России (2011 год). Каждая квартира характеризуется следующими свойствами:

- Количество комнат (1-3);
- Признак этажности (первый/последний или нет);
- Площадь общая, м2;
- Площадь жилая, м2;
- Площадь кухни, м2;
- Наличие агентства – продается объект напрямую или через агентство;
- Состояние квартиры – экспертная оценка по шкале от 2 до 5 (2 – нуждается в ремонте, 5 – отличное состояние квартиры);
- Тип планировки;
- Район – географическая принадлежность;

Результирующий признак – стоимость квартиры в тыс. рублей.

Предварительно проведем аудит выборки при помощи узла «*Качество данных*». Все настройки мастера обработки этого узла оставим предлагаемыми по умолчанию. В результате откроется визуализатор «*Оценка качества данных*».

Аудит данных обнаружил несколько выбросов (выходящих за границы 3-сигма) и экстремальных значений (выходящих за границы 5-сигма). В частности, детализация показывает, что для поля «*Общая площадь*» есть три экстремальных значения 133 и 134 м2 (рис. 10.1).

Вообще, нейросетевые модели достаточно устойчивы к шумам и выбросам, тем не менее, экстремальные значения лучше все-таки удалить. По умолчанию предлагается ограничить найденные выбросы и экстремальные значения.

Переопределим это действие:

- для выбросов выбрать пункт «*Оставить без изменения*»;
- для экстремальных значений – «*Удалять*».

Для того чтобы эти действия были произведены, после узла «*Качество данных*» добавьте узел «*Редактирование выбросов*».

Для оценки качества нейросетевой модели можно использовать прием перекрестной проверки (cross-validation). Это повторение всего процесса обучения и тестирования несколько раз при различных случайных выборках.

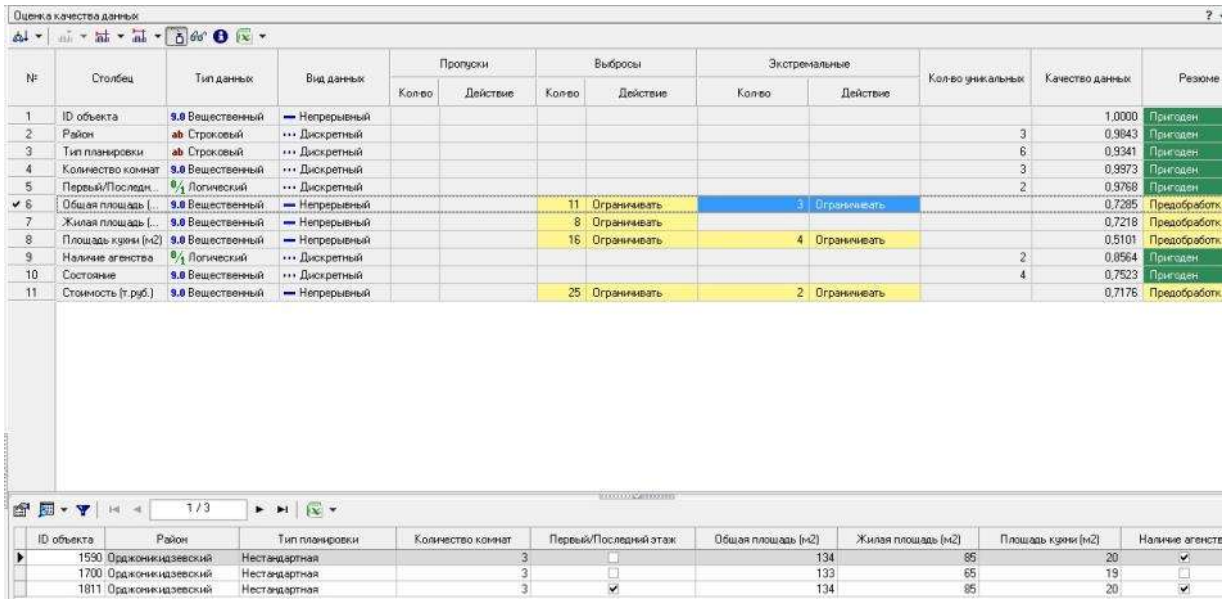


Рис. 10.1

Для определения ошибки принято делать десятиблочную перекрестную проверку. Данные случайным образом разделяются на 10 блоков, в каждом из которых классы наблюдений представлены приблизительно так же, как и в исходном множестве. Затем модель обучается на 9/10 данных и тестируется на оставшейся 1/10 части. Полученные 10 значений ошибки усредняются, и результат рассматривается как общая ошибка модели.

Для того, чтобы заложить эту логику в сценарий необходимо разделить выборку на 10 примерно равных частей. Это делается при помощи нескольких узлов.

а) Узел «*Квантование*» выделяет 10 квантилей, в каждом от 212 до 213 записей (рис. 10.2).

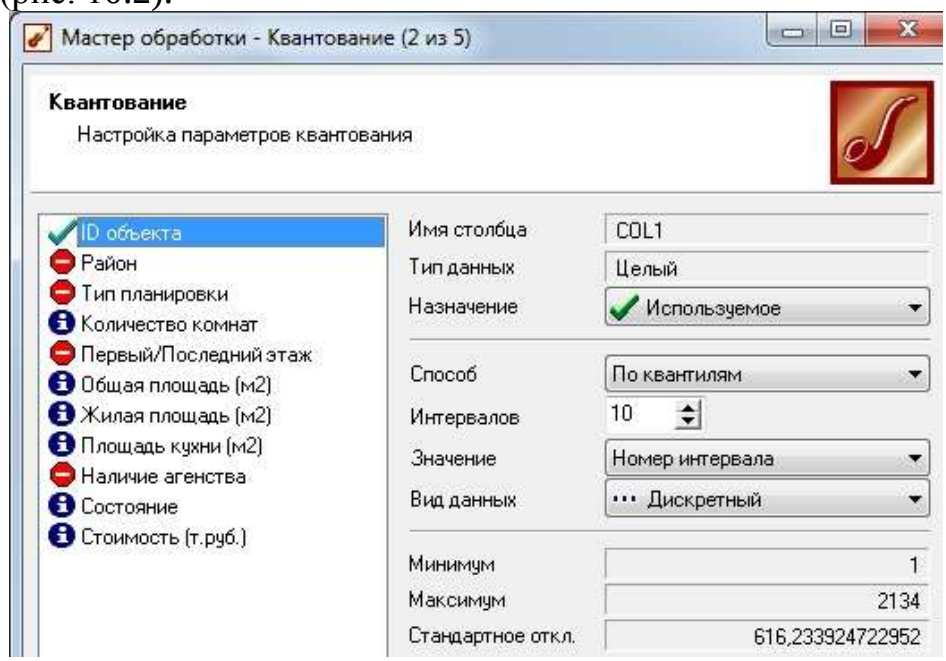


Рис. 10.2

б) Узел «*Группировка*» производит группировку по полю «*ID объекта*» (рис. 10.3).

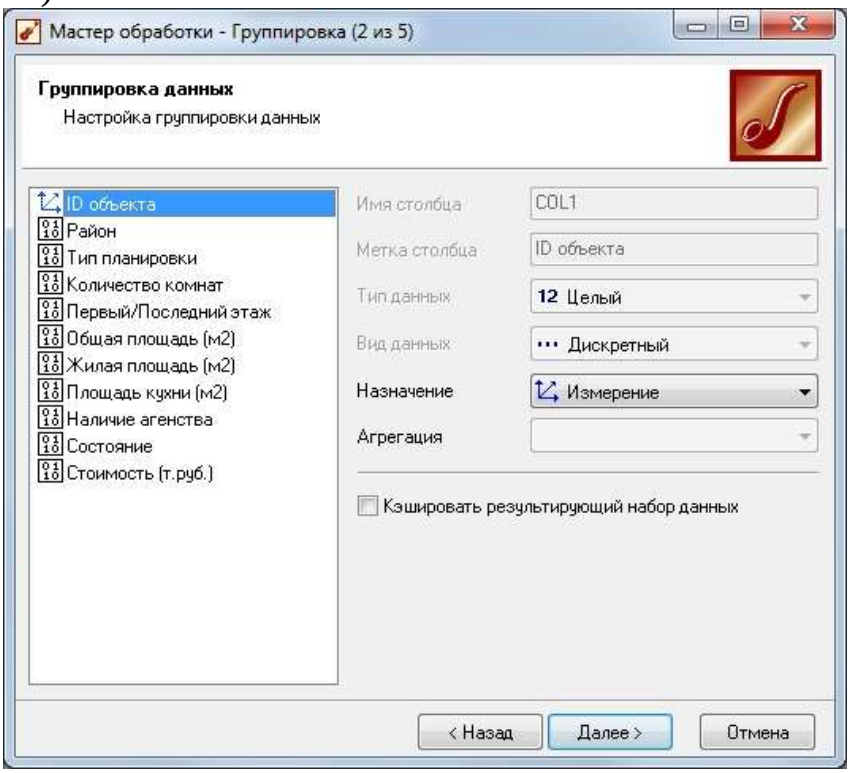


Рис. 10.3

в) Узел «*Настройка набора данных*» формируют список уникальных номеров блоков с меткой № блока и именем Block (рис. 10.4).

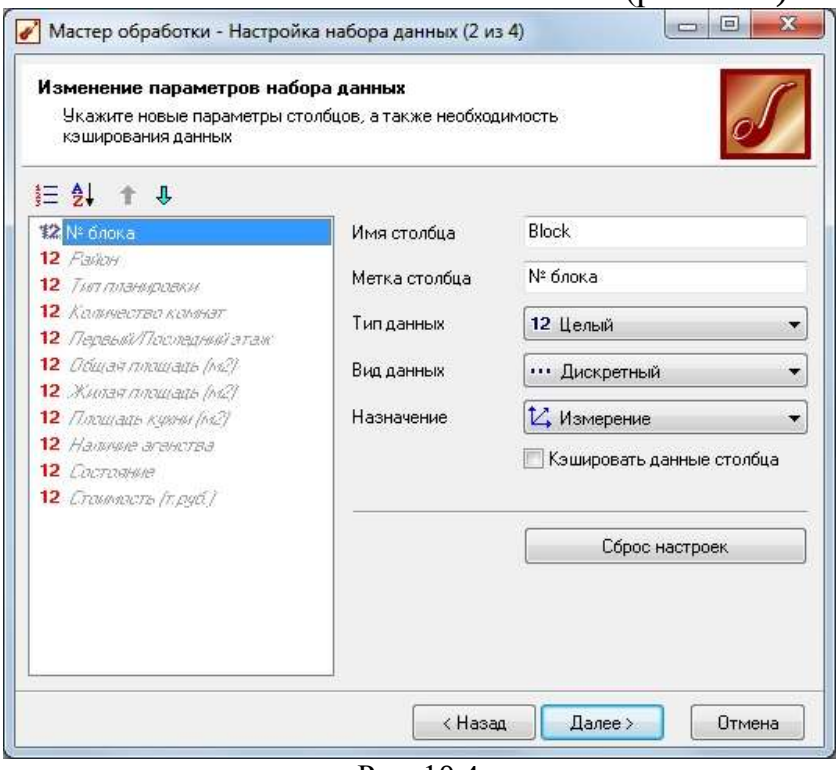


Рис. 10.4

г) Узел «*Слияние с узлом*» (полное внешнее соединение) «размножает» записи исходной выборки (узел «*Квантование*») в число раз, равное коли-

честву блоков – в итоге имеем 21280 записей и идентификатор группы для каждой из них.

Проведем построение нейросети для нулевого блока. Для этого необходимо использовать фильтр.

Выделите тестовое и обучающее множество при помощи «*Калькулятора*», записав в него логическое выражение (рис. 10.5).

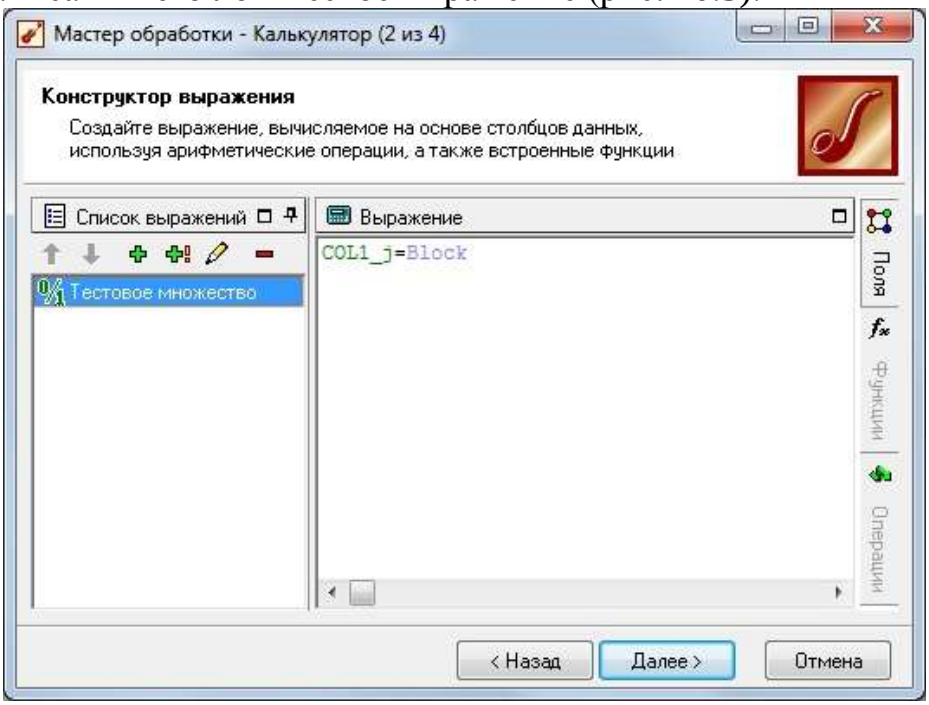


Рис. 10.5

Теперь все готово к построению модели нейросети. Запустите мастер обработки и выберите обработчик «*Нейросеть*» (рис. 10.6).

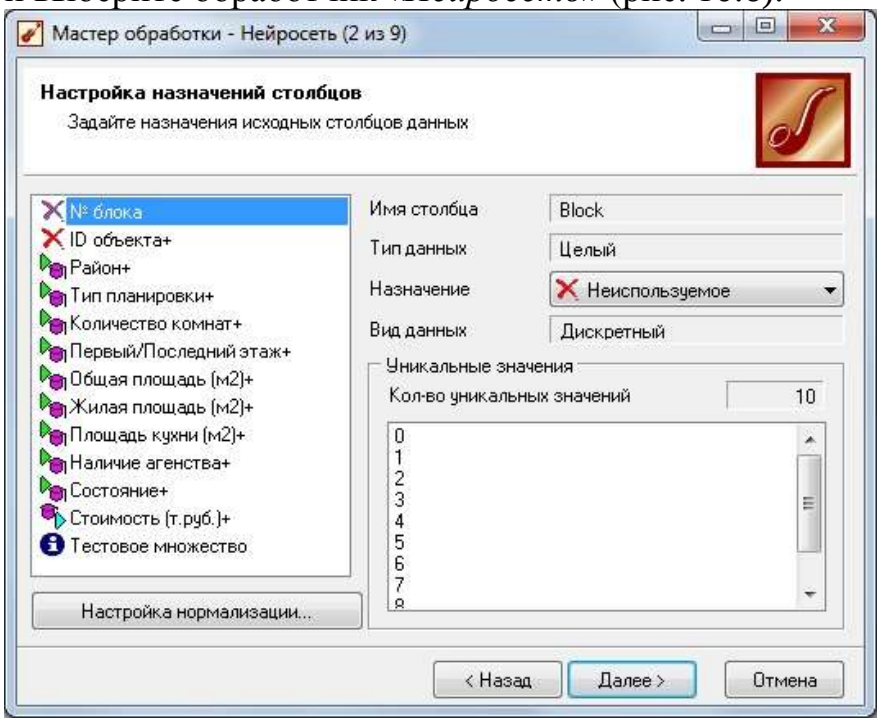


Рис. 10.6

Для полей, содержащих информацию о состоянии, комнатах, этажах и агентстве назначить нормализатор «*Уникальные значения*».

На 3 шаге указать способ разделения – «*по столбцу*» и столбец «*Тестовое множество*».

На 4 шаге настраивается структура нейронной сети. Укажите количество скрытых слоев – 1, а количество нейронов – 5.

На следующих шагах настройки измените только количество эпох, по достижению которых нейросеть останавливает обучение, на 1000. После чего запустите нейросеть на обучение.

Для отображения полученных результатов выберите следующие визуализаторы: «*Граф нейросети»* для отображения структурной схемы построенной нейронной сети; «*Диаграмма рассеяния»* для просмотра качества обучения; «*Что-если»* для расчета стоимости квартиры по введенным пользователям характеристикам.

Рассмотрим визуализатор «*Граф нейросети»* (рис. 10.7). На нем графически отображается нейронная сеть со всеми ее нейронами и синаптическими связями. Значения весов, отображаются определенным цветом, посмотреть которое можно по цветовой шкале, расположенной внизу окна.

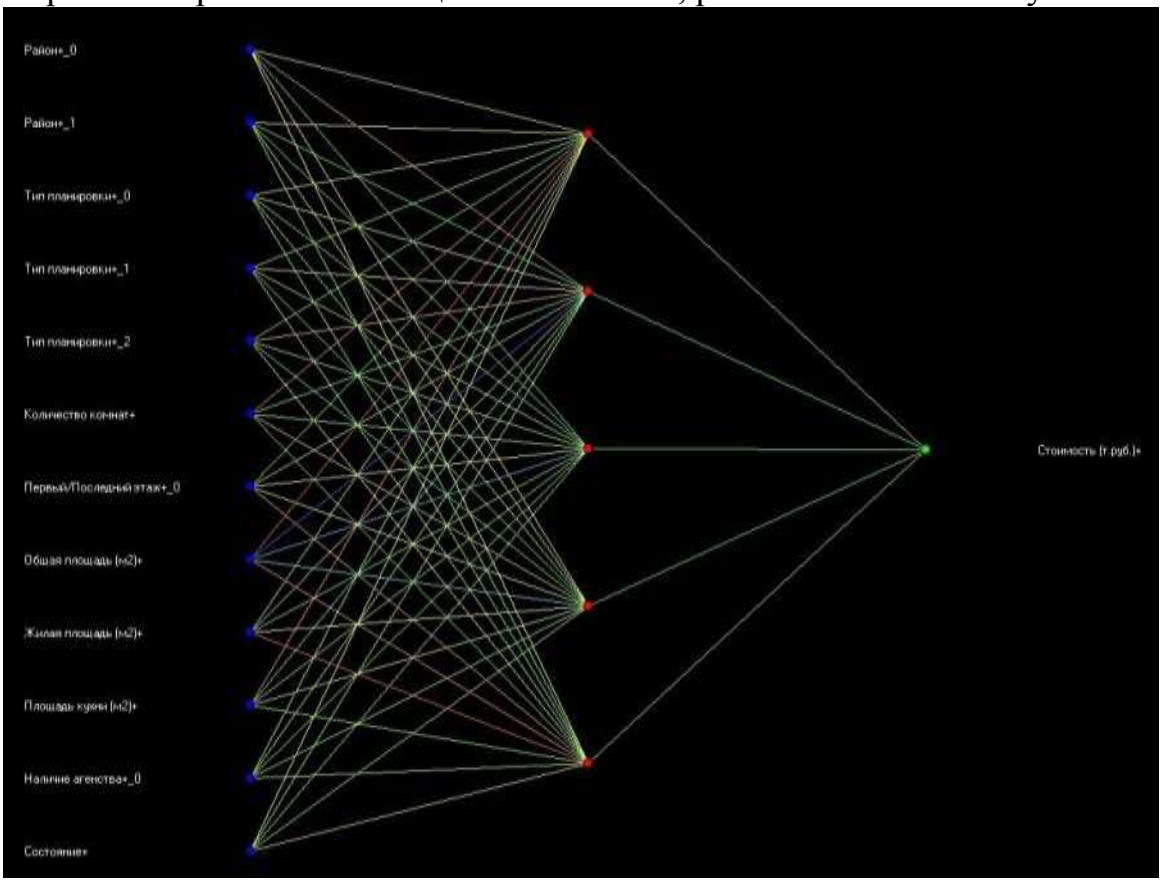


Рис. 10.7

Диаграмма рассеяния показывает качество регрессионной модели. Большая масса точек сосредоточена вблизи линии идеальных значений, поэтому можно сказать, что модель обучилась хорошо (рис. 10.8).

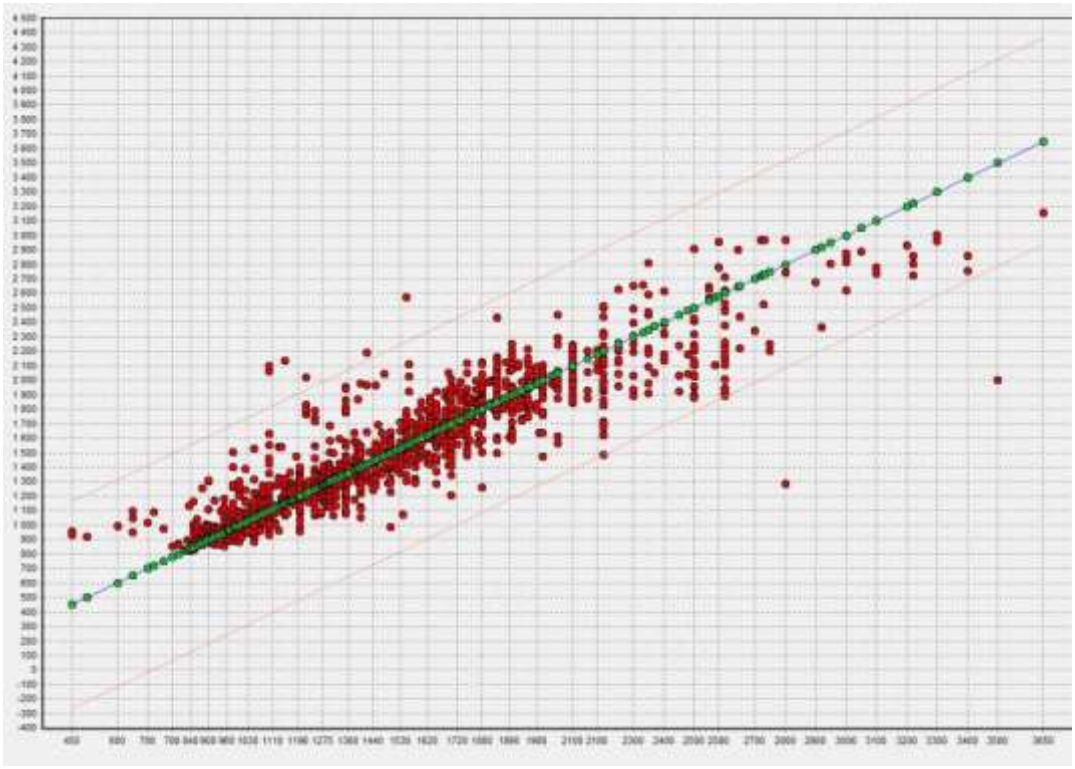


Рис. 10.8

Построение нейросетевой модели для одного блока окончено.

Рассчитаем среднюю ошибку аппроксимации для стоимости недвижимости при помощи калькулятора. Это позволит более точно численно оценить качество модели. Для этого используем «*Калькулятор*» (рис. 10.9). Для расчета количества записей в область *Выражение* ввести 1.

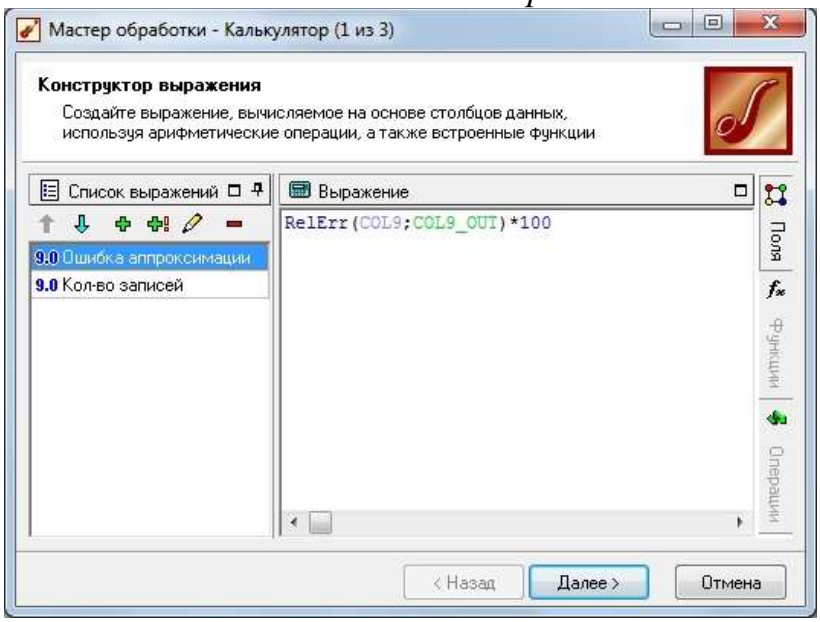


Рис. 10.9

Сгруппируйте данные как показано на рис. 10.10.

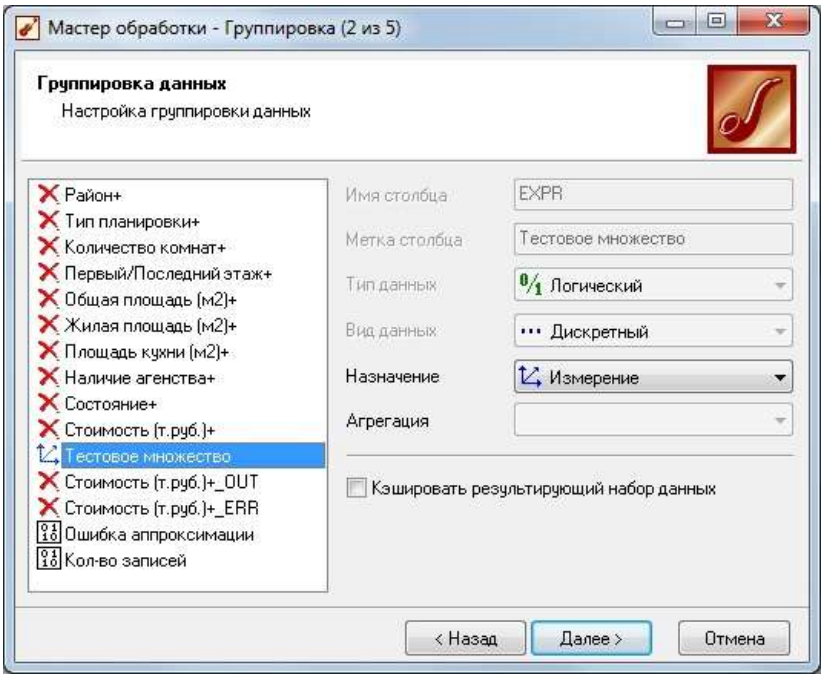


Рис. 10.10

Используя «*Калькулятор»,* добавим новое поле «*Средняя ошибка аппроксимации*», рассчитываемое как отношение ошибки и количества записей. Ошибка получилась в районе 8,0%. Хорошим результатом считается ошибка до 10-12%. Модель является применимой для расчета стоимости недвижимости.

Для проведения 10-блочной кросс-валидации требуется проделать последовательность действий как в предыдущем шаге, но для всех блоков. Это делается при помощи «*Групповой обработки*» от узла «*Внешнее соединение*».

На первом шаге мастера обработки этого узла укажем поле «*№ блока*» как поле, по которому будет проводиться групповая обработка. На следующих двух шагах нужно указать цепочку узлов для групповой обработки. Это будет ветвь от узла фильтра блока до расчета средней ошибки аппроксимации.

В параметрах групповой обработки поставить первый, третий и четвертый флажок.

Запуск групповой обработки всегда приведет к построению 10 моделей нейросетей. В итоге мы получим 10 оценок средней ошибки аппроксимации на обучающем и на тестовом множествах.

Из рис. 10.11 видно, что минимальная ошибка достигается на подвыборке под номером 6. Выберем эту модель как основную и перенастроим ветвь с фильтром на этот номер блока.

№ блока+	Тестовое множество	Ошибка аппроксимации	Кол-во записей	Средняя ошибка
0	☐	14350,6572725895	1915	7,49381580814072
0	☑	1809,98779411943	213	8,49759527755601
1	☐	14352,4520795981	1915	7,49475304417656
1	☑	1816,56910338239	213	8,52849344310981
2	☐	14402,623637272	1916	7,51702695055952
2	☑	1791,24073600041	212	8,44924875471893
3	☐	14604,1516584395	1915	7,62618885558199
3	☑	1855,70648858922	213	8,71223703563016
4	☐	14167,9302536837	1915	7,39839699931261
4	☑	1794,88454917212	213	8,42668802428226
5	☐	14393,0368414795	1915	7,51594613132087
5	☑	1685,02666466472	213	7,91092330828507
6	☐	14590,1838070705	1915	7,618894938418
6	☑	1401,82468404074	213	6,58133654479222
7	☐	14731,1587830082	1916	7,68849623330279
7	☑	1417,28245853265	212	6,68529461572005
8	☐	14463,3383400347	1915	7,55265709662387
8	☑	1654,72420001612	213	7,76865821603813
9	☐	14387,9390302335	1915	7,5132840888948
9	☑	1874,08068440966	213	8,79850086577305

Рис. 10.11

На основе лучшей модели, построенной на подвыборке № 6, спрогнозируем стоимость следующего объекта недвижимости (рис. 10.12). Для этого воспользуемся визуализатором *Что-Если*.

Поле	Значение
Входные	
ab Район	Орджоникидзевский
ab Тип планировки	свердловский вариант
12 Количество комн...	3
0/1 Первый/Послед...	False
9.0 Общая площадь ...	63
9.0 Жилая площадь ...	41
9.0 Площадь кухни (...	8
0/1 Наличие агенства	False
12 Состояние	4
Выходные	
9.0 Стоимость (т.руб.)	1856,4609876801

Рис. 10.12

По прогнозу нейронной сети стоимость квартиры составляет 1856,5 тыс. рублей.

ПРАКТИЧЕСКОЕ ЗАДАНИЕ №1. Моделирование результатов деятельности агрофирмы

Задание на выполнение работы.

Провести анализ результатов деятельности агрофирмы, используя различные методы Data Mining.

Для моделирования использовать файл *agro1.txt*

Вариант 1 (№ зачетной книжки заканчивается на 0-3)

1. Выявить какой из факторов в большей степени влияет на валовый сбор зерна.
2. Построить модель зависимости валового сбора зерна от посевной площади и урожайности.
3. Спрогнозировать валовый сбор зерна при посевной площади 30 000 га и урожайности 30 ц/га.
4. Построить график зависимости валового сбора от посевной площади.
5. Привести значение средней ошибки модели.
6. Построить нейросетевую модель зависимости прибыли, получаемой предприятием, от всех значимых факторов.
7. Построить и проанализировать графики зависимости прибыли от всех использованных факторов.
8. Привести значение средней ошибки модели.

Вариант 2 (№ зачетной книжки заканчивается на 4-7)

1. Выявить какой из факторов в большей степени влияет на объем продаж зерна.
2. Построить модель зависимости реализации зерна от посевной площади и валового сбора.
3. Спрогнозировать объем продаж зерна при посевной площади 30 000 га и валовом сборе 200 000 т.
4. Построить график зависимости объема продаж от валового сбора.
5. Привести значение средней ошибки модели.
6. Построить нейросетевую модель зависимости прибыли, получаемой предприятием, от всех значимых факторов.
7. Построить и проанализировать графики зависимости прибыли от всех использованных факторов.
8. Привести значение средней ошибки модели.

Вариант 3 (№ зачетной книжки заканчивается на 8-9)

1. Выявить какой из факторов в большей степени влияет на прибыль.
2. Построить модель зависимости прибыли от посевной площади и валового сбора
3. Спрогнозировать прибыль от продажи зерна при посевной площади 30 000 га и валового сбора 200 000 т.
4. Построить график зависимости прибыли от валового сбора.
5. Привести значение средней ошибки модели.
6. Построить нейросетевую модель зависимости прибыли, получаемой предприятием, от всех значимых факторов.
7. Построить и проанализировать графики зависимости прибыли от всех использованных факторов.
8. Привести значение средней ошибки модели.

ПРАКТИЧЕСКОЕ ЗАДАНИЕ №2. Анализ сельскохозяйственных предприятий

Задание на выполнение работы.

Провести анализ 100 крупных и эффективных с/х предприятий РФ, используя различные методы Data Mining.

Для моделирования использовать файл *agro2.txt*

Вариант 1 (№ зачетной книжки заканчивается на 0-2)

1. Построить модель, связывающую площадь с/х угодий и численность работников.
2. Спрогнозировать требуемую численность работников для обработки 50 000 га.
3. Провести сегментацию всех предприятий, разбив их на кластеры. Дать характеристику каждому кластеру. Выявить в какие кластеры попали предприятия Ленинградской области.
4. Выявить суммарное число работников предприятий Ленинградской области и средний размер с/х угодий.

Вариант 2 (№ зачетной книжки заканчивается на 3-5)

1. Построить модель, связывающую численность работников и площадь с/х угодий.
2. Спрогнозировать площадь с/х угодий, которую смогут обработать 2000 работников.

3. Провести сегментацию всех предприятий, разбив их на кластеры. Дать характеристику каждому кластеру. Выявить в какие кластеры попали предприятия Краснодарского края.
4. Выявить суммарное число работников предприятий Краснодарского края и средний размер с/х угодий.

Вариант 3 (№ зачетной книжки заканчивается на 6-7)

1. Построить модель, связывающую площадь с/х угодий и численность работников.
2. Спрогнозировать требуемую численность работников для обработки 60 000 га.
3. Провести сегментацию всех предприятий, разбив их на кластеры. Дать характеристику каждому кластеру. Выявить в какие кластеры попали предприятия Ставропольского края.
4. Выявить суммарное число работников предприятий Ставропольского края и средний размер с/х угодий.

Вариант 4 (№ зачетной книжки заканчивается на 8-9)

1. Построить модель, связывающую численность работников и площадь с/х угодий.
2. Спрогнозировать площадь с/х угодий, которую смогут обработать 3000 работников.
3. Провести сегментацию всех предприятий, разбив их на кластеры. Дать характеристику каждому кластеру. Выявить в какие кластеры попали предприятия Московской области.
4. Выявить суммарное число работников предприятий Московской области и средний размер с/х угодий.

ПРАКТИЧЕСКОЕ ЗАДАНИЕ №3. Анализ энергетической отрасли США

Задание на выполнение работы.

Провести анализ и прогнозирование продаж сегментов энергетической отрасли по штатам США, используя различные методы Data Mining.

Для моделирования использовать файл *sales_annual.txt*

Вариант 1 (№ зачетной книжки заканчивается на 0-2)

1. Построить прогнозную модель суммарных продаж энергии по штату Нью-Йорк (NY) на период с 2015 по 2020 годы.

2. Провести сегментацию всех штатов. Дать характеристику каждому кластеру. Выявить в какой кластер попал штат Нью-Йорк.

Вариант 2 (№ зачетной книжки заканчивается на 3-5)

1. Построить прогнозную модель суммарных продаж энергии по штату Аляска (AL) на период с 2015 по 2020 годы.
2. Провести сегментацию всех штатов. Дать характеристику каждому кластеру. Выявить в какой кластер попал штат Аляска.

Вариант 3 (№ зачетной книжки заканчивается на 6-9)

1. Построить прогнозную модель суммарных продаж энергии по штату Техас (TX) на период с 2015 по 2020 годы.
2. Провести сегментацию всех штатов. Дать характеристику каждому кластеру. Выявить в какой кластер попал штат Техас.

ПРАКТИЧЕСКОЕ ЗАДАНИЕ №4. Анализ розничных продаж

Задание на выполнение работы.

Провести анализ продаж чистящих средств, используя различные методы Data Mining.

Для моделирования использовать файл *sample-sales-data.txt*

Вариант 1 (№ зачетной книжки заканчивается на 0-2)

1. Выявить от чего зависит выбор канала покупки покупателем.
2. Выявить какой регион приносит компании больший доход.
3. Провести сегментацию клиентов и выявить, кто приносит компании больший доход.

Вариант 2 (№ зачетной книжки заканчивается на 3-5)

1. Выявить от чего зависит выбор канала покупки покупателем.
2. Выявить какой способ продажи приносит компании больший доход.
3. Провести сегментацию клиентов и выявить, кто приносит компании больший доход.

Вариант 3 (№ зачетной книжки заканчивается на 6-9)

1. Выявить от чего зависит выбор канала покупки покупателем.
2. Выявить какой продукт приносит компании больший доход.

3. Провести сегментацию клиентов и выявить, кто приносит компании больший доход.

ПРАКТИЧЕСКОЕ ЗАДАНИЕ №5. Анализ совершенных преступлений

Задание на выполнение работы.

Провести анализ совершенных преступлений в районах Лос-Анджелеса, используя различные методы Data Mining.

Для моделирования использовать файл *crimes.txt*

Вариант 1 (№ зачетной книжки заканчивается на 0-2)

1. Выявить в каком районе чаще всего совершаются преступления.
2. Выявить от чего зависит факт раскрытия преступления.
3. Определить будет ли раскрыто убийство женщины 40 лет в квартире Голливуде.

Вариант 2 (№ зачетной книжки заканчивается на 3-5)

1. Выявить самое часто встречающееся место преступления.
2. Выявить от чего зависит факт раскрытия преступления.
3. Определить будет ли раскрыто похищение 10-летнего мальчика на улице в Западном Лос-Анджелесе.

Вариант 3 (№ зачетной книжки заканчивается на 6-9)

1. Выявить самое распространенное преступление.
2. Выявить от чего зависит факт раскрытия преступления.
3. Определить будет ли раскрыта кража в метро в районе Уилшир у мужчины 80 лет.

ПРАКТИЧЕСКОЕ ЗАДАНИЕ №6. Анализ спортивных результатов

Задание на выполнение работы.

Провести анализ результатов кросса, используя различные методы Data Mining.

Для моделирования использовать файл *cross.txt*

Вариант 1 (№ зачетной книжки заканчивается на 0-2)

1. Спрогнозировать результат мужчины 1980 года рождения из Санкт-Петербурга двумя разными способами (нейронная сеть и линейная регрессия).

2. Выявить от каких факторов в большей степени зависит результат.
3. Рассчитать ошибку прогнозирования по каждому методу.
4. Провести сегментирование спортсменов и проанализировать каждый кластер.

Вариант 2 (№ зачетной книжки заканчивается на 3-5)

1. Спрогнозировать результат мужчины 1990 года рождения из Самары двумя разными способами (нейронная сеть и линейная регрессия).
2. Выявить от каких факторов в большей степени зависит результат.
3. Рассчитать ошибку прогнозирования по каждому методу.
4. Провести сегментирование спортсменов и проанализировать каждый кластер.

Вариант 3 (№ зачетной книжки заканчивается на 6-9)

1. Спрогнозировать результат женщины 1975 года рождения из Москвы двумя разными способами (нейронная сеть и линейная регрессия).
2. Выявить от каких факторов в большей степени зависит результат.
3. Рассчитать ошибку прогнозирования по каждому методу.
4. Провести сегментирование спортсменов и проанализировать каждый кластер.