

**Задание на расчетно-графическое задание**

по дисциплине **Интеллектуальный анализ данных**

**Тема: Компьютерные методы анализа данных и прогнозирования.**

**Цель:** изучить основы методов анализа экспериментальных данных и освоить технику их практического применения в Deductor Studio

**Порядок выполнения**

**Задание 1. Решение задачи поиска ассоциаций в Deductor**

- 1.1. Подготовить данные для поиска ассоциативных данных.
- 1.2. Провести поиск ассоциативных правил, проанализировать полученные результаты

**Задание 2. Прогнозирование временного ряда.**

- 2.1. Подготовить данные для прогнозирования временного ряда
- 2.2. Провести анализ временного ряда
- 2.3. Провести прогнозирование временного ряда на 12 месяцев.

**Список литературы**

1. Deductor. Руководство аналитика. Версия 5.3
2. Кацко И.А., Паклин Н.Б. Практикум по анализу данных на компьютере / Под ред. Г.В. Гореловой. – М.: КолосС. 2009. - 2/8 с.: ил. – (Учебники и учеб, пособия для студентов высш. учеб, заведений).

## 1. Решение задачи поиска ассоциаций в Deductor

### 1.1. Краткие теоретические сведения

Ассоциативные правила позволяют находить закономерности между связанными событиями. Примером такого правила служит утверждение, что покупатель, приобретающий "Хлеб", приобретет и "Молоко". Впервые эта задача была предложена для поиска ассоциативных правил для нахождения типичных шаблонов покупок, совершаемых в супермаркетах, поэтому иногда ее еще называют анализом рыночной корзины (market basket analysis)

Транзакция – это множество событий, произошедших одновременно. Пусть имеется база данных, состоящая из покупательских транзакций. Каждая транзакция – это набор товаров, купленных покупателем за один визит.

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 возможность находить очень простые и интуитивно понятные правила.

Основными характеристиками таких правил являются **поддержка** и **достоверность**. Правило "Из  $X$  следует  $Y$ " имеет поддержку  $s$ , если  $s\%$  транзакций из всего набора содержат наборы элементов  $X$  и  $Y$ . Достоверность правила показывает, какова вероятность того, что из  $X$  следует  $Y$ . Правило "Из  $X$  следует  $Y$ " справедливо с достоверностью  $c$ , если  $c\%$  транзакций из всего множества, содержащих набор элементов  $X$ , также содержат набор элементов  $Y$ .

Пример: пусть 75% транзакций, содержащих хлеб, также содержат молоко, а 3% от общего числа всех транзакций содержат оба товара. 75% – это достоверность правила, а 3% – это поддержка.

Лифт – это отношение частоты появления условия в транзакциях, которые также содержат и следствие, к частоте появления следствия в целом. Значения лифта, большие единицы, показывают, что условие появляется более часто в транзакциях, содержащих и следствие, чем в остальных.

Алгоритмы поиска ассоциативных правил предназначены для нахождения всех правил вида "из  $X$  следует  $Y$ ", причем поддержка и достоверность этих правил должны находиться в рамках некоторых наперед заданных границ, называемых соответственно минимальной и максимальной поддержкой и минимальной и максимальной достоверностью. Границы значений параметров поддержки и достоверности выбираются таким образом, чтобы ограничить количество найденных правил. Если поддержка имеет большое значение, то алгоритмы будут находить правила, хорошо известные аналитикам или настолько очевидные, что нет никакого смысла проводить такой анализ. С другой стороны, низкое значение поддержки ведет к генерации огромного количества правил, что, конечно, требует существенных вычислительных ресурсов. Тем не менее, большинство

интересных правил находится именно при низком значении порога поддержки, хотя слишком низкое значение поддержки ведет к генерации статистически необоснованных правил. Таким образом, необходимо найти компромисс, обеспечивающий, во-первых, интересность правил и, во-вторых, их статистическую обоснованность. Поэтому значения этих границ напрямую зависят от характера анализируемых данных и подбираются индивидуально. Еще одним параметром, ограничивающим Количество найденных правил является максимальная мощность часто встречающихся множеств. Если этот параметр указан, то при поиске правил будут рассматриваться только множества, количество элементов которых будет не больше данного параметра.

Обычные ассоциативные правила – это правила, в которых как в условии, так и в следствии присутствуют только элементы транзакций и при вычислении которых используется только информация о том, присутствует ли элемент в транзакции или нет.

Все множество ассоциативных правил можно разделить на три вида:

Полезные правила – содержат действительную информацию, которая ранее была неизвестна, но имеет логичное объяснение. Такие правила могут быть использованы для принятия решений, приносящих выгоду.

Тривиальные правила – содержат действительную и легко объяснимую информацию, которая уже известна. Такие правила, хотя и объяснимы, но не могут принести какой-либо пользы, т.к. отражают или известные законы в исследуемой области, или результаты прошлой деятельности. При анализе рыночных корзин в правилах с самой высокой поддержкой и достоверностью окажутся товары-лидеры продаж. Практическая ценность таких правил крайне низка.

Непонятные правила – содержат информацию, которая не может быть объяснена. Такие правила могут быть получены или на основе аномальных значений, или глубоко скрытых знаний. Напрямую такие правила нельзя использовать для принятия решений, т.к. их необъяснимость может привести к непредсказуемым результатам. Для лучшего понимания требуется дополнительный анализ.

## **1.2. Поиск ассоциативных правил в Deductor Studio**

Для поиска обычных ассоциативных правил в программе служит обработчик «Ассоциативные правила».

Обработчик требует на входе два поля: идентификатор транзакции и элемент транзакции. Например, идентификатор транзакции – это номер чека или код клиента. А элемент – это наименование товара в чеке или услуга, заказанная клиентом.

Оба поля (идентификатор и элемент транзакции) должны быть дискретного вида.

Затем следует настройка параметров поиска правил. Всего четыре параметра:

**Минимальная и максимальная поддержка.** Ассоциативные правила ищутся только в некотором множестве всех транзакций. Для того чтобы транзакция вошла в это множество, она должна встретиться в исходной выборке количество раз, больше минимальной поддержки и меньше максимальной. Например, минимальная поддержка равна 1%, а максимальная – 20%. Количество элементов «Хлеб» и «Молоко» столбца «Товар» с одинаковым значением столбца «Номер чека» встречаются в 5% всех транзакций (номеров чека). Тогда эти две строки войдут в искомое множество.

**Минимальная и максимальная достоверность.** Это процентное отношение количества транзакций, содержащих все элементы, которые входят в правило, к количеству транзакций, содержащих элементы, которые входят в условие. Если транзакция – это заказ, а элемент – товар, то достоверность характеризует, насколько часто покупаются товары, входящие в следствие, если заказ содержит товары, вошедшие во всё правило.

#### **Пример:**

Рассмотрим механизм поиска ассоциативных правил на примере данных о продажах товаров в некоторой торговой точке. Данные представляются в виде таблицы, в которой представлена информация по покупкам продуктов нескольких групп. Она имеет всего два поля "Номер чека" и "Товар".

N	Поле	Тип поля	Назначение
1	ID	строковый	Код чека
2	ITEM	строковый	Товар

Необходимо решить задачу анализа потребительской корзины с целью последующего применения результатов для стимулирования продаж.

Для поиска ассоциативных правил необходимо запустить **Мастер обработки**. В нем выбрать тип обработки "Ассоциативные правила".

На втором шаге Мастера следует указать, какой столбец является идентификатором транзакции (чек), который должен быть дискретным, а какой элементом транзакции (товар).

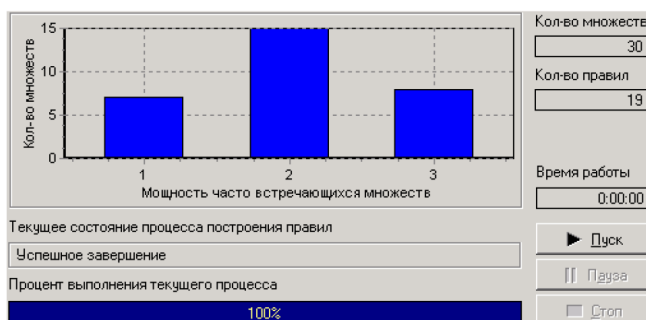
ID	Номер чека	Имя столбца	COL2
	Товар	Тип данных	Строковый
		Назначение	Элемент
		Вид данных	Дискретный
		Уникальные значения Кол-во уникальных значений: 7	
		ВАФЛИ КЕТЧУПЫ, СОУСЫ, АДЖИКА МАКАРОННЫЕ ИЗДЕЛИЯ МЕД СУХАРИ СЫРЫ ЧАЙ	

Следующий шаг позволяет настроить параметры построения ассоциативных правил: минимальную и максимальную поддержку, минимальную и максимальную достоверность, а также максимальную мощность множества. Эти параметры необходимо выставлять исходя из характера имеющихся данных. Например, границы поддержки следует указать – 13% и 80% и достоверности 60% и 90%. Можно оставить по умолчанию.

**Максимальная мощность искомых часто встречающихся множеств** – параметр ограничивает длину k-предметного набора. Например, при установке значения 4 шаг генерации популярных наборов будет остановлен после получения множества 4-предметных наборов. В конечном итоге это позволяет избежать появления длинных ассоциативных правил, которые трудно интерпретируются.

Часто встречающиеся множества	
Минимальная поддержка, %	13
Максимальная поддержка, %	80
<input type="checkbox"/> Максимальная мощность искомых часто встречающихся множеств	4
Ассоциативные правила	
Минимальная достоверность, %	60
Максимальная достоверность, %	90

Следующий шаг позволяет запустить процесс поиска ассоциативных правил. На экране отображается информация о количестве множеств и найденных правил, а также числе часто встречающихся множеств.



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. На сколько часто встречается множество в исходном наборе транзакций, можно судить по поддержке. Данный визуализатор отображает множества в виде списка.

№	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

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. Этот список представлен таблицей со столбцами: "Номер правила", "Условие", "Следствие", "Поддержка, %", "Поддержка, Количество", "Достоверность".

Таким образом, эксперту предоставляется набор правил, которые описывают поведение покупателей.

Правил: 63 из 63      Фильтр: Без фильтрации							
№	Номер правила	Условие	Следствие	Поддержка		Достоверность	Лифт
				Кол-во	%		
1	60	Клей - ж. гвозди	Герметики	2	4.55	40.00	2.933
		Шпатлёвка	Пена монтажная				
2	57	Герметики	Клей - ж. гвозди	2	4.55	33.33	2.933
		Пена монтажная	Шпатлёвка				
3	59	Герметики	Клей - ж. гвозди	2	4.55	40.00	2.514
		Шпатлёвка	Пена монтажная				

Например, в правиле с номером 60 в условии присутствуют два элемента: Клей-ж. гвозди и Шпатлёвка. Это правило показывает, что человек, купивший Клей-ж.гвозди и Шпатлёвка с вероятностью 40% купит ещё и Герметики и Пену монтажная.

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.

Ассоциативные правила (по у		Количество правил: 4; Условие: Клей - ж. гвозди			
<ul style="list-style-type: none"> <li>Герметики (31.82%; 14)</li> <li>Клей - ж. гвозди (31.82%; 10) <ul style="list-style-type: none"> <li>Герметики (22.73%; 10)</li> <li>Пена монтажная (15.9%)</li> <li>Шпатлёвка (11.36%; 5)</li> <li>Герметики И Пена монтажная (9.09%; 4)</li> </ul> </li> <li>Пена монтажная (50.00%; 1)</li> </ul>	Следствие	Поддержка		Достоверность, %	Лифт
		Кол-во	%		
	Герметики	10	22.70	71.40	2.245
	Пена монтажная	7	15.90	50.00	1
	Шпатлёвка	5	11.40	35.70	0.827
	Герметики И Пена монтаж...	4	9.09	28.60	2.095

Узлы Герметики, Клей-ж. гвозди, Пена монтажная находятся на верхнем уровне дерева и представляют собой условия. А Герметики, Пена монтажная, Шпатлёвка и т.д. – следствия. Это означает, что человек, купивший Клей-ж. гвозди, купит еще и Герметик с достоверностью 71,40%, пену монтажную с достоверностью 50,00% и т.д. В окне слева расположен список со следствиями для конкретного узла с условием. Для каждого следствия указана поддержка, достоверность и лифт. Например, в исходной выборке данных герметики встретились в 10 транзакциях (чеках).

Второй вариант дерева правил – дерево, построенное по следствию. Здесь на первом уровне располагаются узлы со следствием.

Ассоциативные правила (по с		Количество правил: 8; Следствие: Герметики			
<ul style="list-style-type: none"> <li>Клей - ж. гвозди (31.82%; 10)</li> <li>Герметики (31.82%; 14)</li> <li>Клей - ж. гвозди (22.73%; 10)</li> <li>Пена монтажная (13.6%; 5)</li> <li>Шпатлёвка (11.36%; 5)</li> <li>Клей - ж. гвозди И Пена монтажная (9.09%; 4)</li> <li>Клей - ж. гвозди И Шпатлёвка (6.82%; 3)</li> <li>Клей - ж. гвозди И Эмали (2.27%; 1)</li> <li>Пена монтажная И Шпатлёвка (3.64%; 2)</li> <li>Клей - ж. гвозди И Пена монтажная (2.27%; 1)</li> </ul>	Условие	Поддержка		Достоверность, %	Лифт
		Кол-во	%		
	Клей - ж. гвозди	10	22.70	71.40	2.245
	Пена монтажная	6	13.60	27.30	0.857
	Шпатлёвка	5	11.40	26.30	0.827
	Клей - ж. гвозди И Пена монтаж...	4	9.09	57.10	1.796
	Клей - ж. гвозди И Шпатлёв...	3	6.82	60.00	1.886
	Клей - ж. гвозди И Эмали	1	2.27	50.00	1.571
	Пена монтажная И Шпатлёв...	3	6.82	37.50	1.179
	Клей - ж. гвозди И Пена монтаж...	2	4.55	66.70	2.095

Например, для того чтобы человек приобрел Герметик, он должен купить хотя бы один предмет из следующего списка: Клей-ж.гвозди, Пена монтажная, Шпатлёвка и т.д. И для каждого из этих правил отображены поддержка, достоверность и лифт.

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. Данный визуализатор отображает те же самые правила, что и предыдущий, но в более удобной для анализа форме.

Анализ "Что-если" в ассоциативных правилах 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". 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.

Элемент	Поддержка, %
ВАФЛИ	31.82
КЕТЧУПЫ, СОУСЫ...	52.27
МАКАРОННЫЕ ИЗД...	54.55
МЕД	50.00
СУХАРИ	31.82
СЫРЫ	43.18
ЧАЙ	75.00

Условие	Элемент
	ВАФЛИ
	МЕД

Количество правил: 3

Следствие	Поддержка	
	№	%
ЧАЙ	18	40.90
СУХАРИ	10	22.70
СУХАРИ И ЧАЙ	9	20.50

Результаты анализа можно применить и для сегментации покупателей по поведению при покупках, и для анализа предпочтений клиентов, и для планирования расположения товаров в супермаркетах, кросс-маркетинге. Предлагаемый набор визуализаторов позволяет эксперту найти интересные, необычные закономерности, понять, почему так происходит, и применить их на практике.

### 1.3. Решение задачи поиска ассоциаций в Deductor

Подготовить данные для поиска ассоциативных данных в формате \*.txt.

Провести поиск ассоциативных правил, проанализировать полученные результаты

## 2. Прогнозирование временных рядов

### 2.1. Краткие теоретические сведения

**Временным рядом** (рядом динамики, динамическим рядом) называется последовательность значений показателя или признака, упорядоченная в хронологическом порядке, т.е. в порядке возрастания временного параметра. Отдельные наблюдения временного ряда называются **уровнями** этого ряда.

Каждый временной ряд содержит два элемента:



- 1) значения времени;
- 2) соответствующие им значения уровней ряда.

В качестве показателя времени в рядах динамики могут указываться либо определенные моменты времени (даты), либо отдельные периоды (сутки, месяцы, кварталы, полугодия, годы и т.д.). В зависимости от характера временного параметра ряды делятся на моментные и интервальные. В моментных рядах динамики уровни характеризуют значения показателя по состоянию на определенные моменты времени. Например, моментными являются временные ряды цен на определенные виды товаров, ряды курсов акций, уровни которых фиксируются для конкретных чисел. Примерами моментных рядов динамики могут служить также ряды численности населения или стоимости основных фондов, т.к. значения уровней этих рядов определяются ежегодно на одно и то же число. В интервальных рядах уровни характеризуют значение показателя за определенные интервалы (периоды) времени. Примерами могут служить ряды годовой (месячной, квартальной) динамики производства продукции в натуральном или стоимостном выражении.

Если уровни ряда представляют собой не непосредственно наблюдаемые значения, а производные величины (средние или относительные), то такие ряды называются производными. Уровни этих временных рядов получаются с помощью некоторых вычислений на основе абсолютных показателей.

Для успешного изучения динамики процесса важно, чтобы информация была полной, временной ряд имел достаточную длину (с учетом конкретных целей исследования). Например, при изучении периодических колебаний желательно иметь информацию не менее чем за три полных периода колебания. Поэтому при анализе сезонных колебаний на базе рядов месячной или квартальной динамики желательно иметь информацию, как правило, не менее чем за 3 года. Применение определенного математического аппарата также накладывает ограничение на допустимую длину временных рядов. Например, для использования регрессионного анализа требуется иметь временные ряды, длина которых в несколько раз превосходит количество независимых переменных. Уровни рядов динамики могут содержать аномальные значения или «выбросы». Часто появление таких значений может быть вызвано ошибками при сборе, записи и передаче информации. Выявление и исключение таких значений, замена их истинными или расчетными является необходимым этапом первичной обработки данных, т.к. применение математических методов к «засоренной» информации приводит к искажению результатов анализа. Однако аномальные значения могут отражать реальное развитие процесса, как, например, «скачок» курса доллара в «черный вторник».

В практике исследования динамики явлений и прогнозирования принято считать, что значения уровней временных рядов экономических показателей могут содержать следующие компоненты (составные части или структурно-образующие элементы):

- тренд;
- сезонную компоненту;
- циклическую компоненту;
- случайную составляющую.

Под **трендом** понимают изменение, определяющее общее направление развития, основную тенденцию временного ряда. Это систематическая составляющая долговременного действия. Наряду с долговременными тенденциями во временных рядах экономических процессов часто имеют место более или менее регулярные колебания – периодические составляющие рядов динамики. Если период колебаний не превышает одного года, то их называют **сезонными**. Чаще всего причиной их возникновения считаются природно-климатические условия. Примером могут служить колебания цен на сельскохозяйственную продукцию, в частности на картофель.

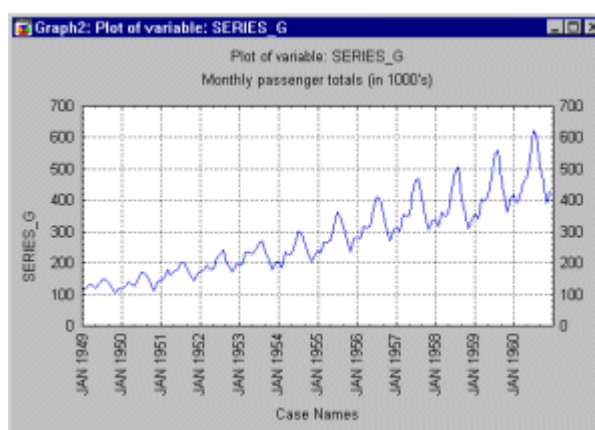
При большем периоде колебания считают, что во временных рядах имеет место **циклическая** составляющая. Примерами могут служить демографические, инвестиционные и другие циклы. Если из временного ряда удалить тренд и периодические составляющие, то останется **нерегулярная** компонента.

Разделяют факторы, под действием которых формируется нерегулярная компонента, на 2 вида:

- факторы резкого, внезапного действия;
- текущие факторы.

Факторы первого вида (например, стихийные бедствия, эпидемии и др.), как правило, вызывают более значительные отклонения. Иногда такие отклонения называют катастрофическими колебаниями. Факторы второго вида вызывают случайные колебания, являющиеся результатом действия большого числа побочных причин. Влияние каждого из текущих факторов незначительно, но ощущается их суммарное воздействие.

**Пример.**



Ряд G представляет месячные международные авиаперевозки (в тысячах) в течение 12 лет с 1949 по 1960. График месячных перевозок ясно показывает почти линейный тренд, т.е. имеется устойчивый рост перевозок из года в год (примерно в 4 раза больше пассажиров перевезено в 1960 году, чем в 1949). В то же время характер месячных перевозок повторяется, они

имеют почти один и тот же характер в каждом годовом периоде (например, перевозок больше в отпускные периоды, чем в другие месяцы). Этот пример показывает довольно определенный тип модели временного ряда, в которой амплитуда сезонных изменений увеличивается вместе с трендом.

## 2.2. Прогнозирование в Deductor Studio

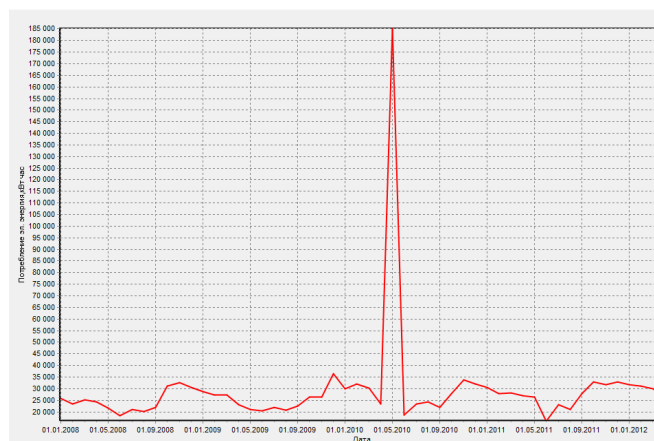
Прогнозирование результата на определенное время вперед, основываясь на данных за прошедшее время, – задача, встречающаяся довольно часто. К примеру, перед большинством торговых фирм стоит задача оптимизации складских запасов, для решения которой требуется знать, что и сколько должно быть продано через 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).

### Пример.

Пример исходных данных по потреблению электроэнергии находятся в Приложении А.

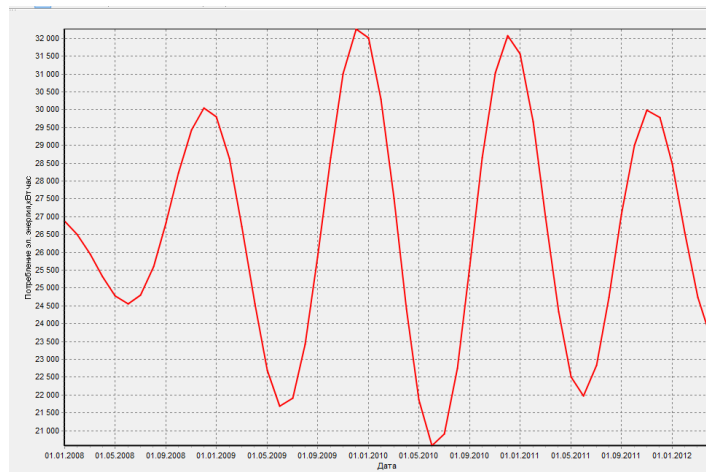
После импорта данных воспользуемся диаграммой для их просмотра.



На ней видно, что данные содержат аномалии (выбросы) и шумы, за которыми трудно разглядеть тенденцию. Поэтому перед прогнозированием необходимо удалить аномалии и сгладить данные.

Следующим шагом необходимо провести «Редактирование выбросов и экстремальных значений» и «Спектральную обработку»

Для просмотра данных можно воспользоваться диаграммой. Видно, что данные сгладились, аномалии и шумы исчезли.



Строить прогноз на будущее будем, основываясь на данных прошлых периодов, т. е. предполагая, что количество потребляемой энергии на следующий месяц зависит от количества потребляемой энергии за предыдущие месяцы. Это значит, что входными факторами для модели могут быть количество потребляемой энергии за текущий месяц, продажи за месяц ранее и т.д., а результатом должно быть количество потребляемой энергии за следующий месяц.

Для дальнейшей работы необходимо трансформировать данные к скользящему окну.

### Скользящее окно

При решении некоторых задач, например, при прогнозировании временных рядов при помощи нейросети, требуется подавать на вход модели значения нескольких смежных отсчетов из исходного набора данных. Такой метод отбора данных называется скользящим окном (окно – поскольку выделяется только некоторый непрерывный участок данных, скользящее – поскольку это окно «перемещается» по всему набору). При этом эффективность реализации заметно повышается, если не выбирать данные каждый раз из нескольких последовательных записей, а последовательно расположить данные, относящиеся к конкретной позиции окна, в одной записи.

Значения в одном из полей записи будут относиться к текущему отсчету, а в других – смещены от текущего отсчета «в будущее» или «в прошлое». Таким образом, преобразование скользящего окна имеет два параметра: «глубина погружения» - количество «прошлых» отсчетов, попадающих в окно, и «горизонт прогнозирования» – количество «будущих» отсчетов. Следует отметить, что для граничных (относительно начала и конца всей выборки) положений окна будут формироваться неполные записи, т.е. записи, содержащие пустые значения для отсутствующих прошлых или будущих отсчетов. Алгоритм преобразования позволяет

исключить такие записи из выборки (тогда для нескольких граничных отсчетов записи формироваться не будут) либо включить их (тогда формируются записи для всех имеющихся отсчетов, но некоторые из них будут неполными). Для правильного формирования скользящего окна данные должны быть соответствующим образом упорядочены.

### Скользящее окно 12 месяцев назад

Запустить Мастер обработки, выбрать в качестве обработчика скользящее окно и перейти на следующий шаг.

Требуется выбрать глубину погружения 12, назначив поле "Количество" используемым. Тогда данные трансформируются к скользящему окну так, что аналитику будут доступны все нужные факторы для построения прогноза.

Дата	Потребление эл. энергии кВт час-12	Потребление эл. энергии кВт час-11	Потребление эл. энергии кВт час-10	Потребление эл. энергии кВт час-9
01.01.2009	26879,7916826241	26497,3238742571	25947,0310252638	25321,9720601086
01.02.2009	26497,3238742571	25947,0310252638	25321,9720601086	24792,945447299
01.03.2009	25947,0310252638	25321,9720601086	24792,945447299	24566,112737449
01.04.2009	25321,9720601086	24792,945447299	24566,112737449	24813,6544685508
01.05.2009	24792,945447299	24566,112737449	24813,6544685508	25599,24103298
01.06.2009	24566,112737449	24813,6544685508	25599,24103298	26826,0695167917
01.07.2009	24813,6544685508	25599,24103298	26826,0695167917	28231,4799139563
01.08.2009	25599,24103298	26826,0695167917	28231,4799139563	29438,9021035691
01.09.2009	26826,0695167917	28231,4799139563	29438,9021035691	30058,817094141
01.10.2009	28231,4799139563	29438,9021035691	30058,817094141	29811,7693287039
01.11.2009	29438,9021035691	30058,817094141	29811,7693287039	28634,9592824178
01.12.2009	30058,817094141	29811,7693287039	28634,9592824178	26734,4315637277
01.01.2010	29811,7693287039	28634,9592824178	26734,4315637277	24558,4655310325
01.02.2010	28634,9592824178	26734,4315637277	24558,4655310325	22691,1360235333
01.03.2010	26734,4315637277	24558,4655310325	22691,1360235333	21691,1459590879
01.04.2010	24558,4655310325	22691,1360235333	21691,1459590879	21921,4007057107
01.05.2010	22691,1360235333	21691,1459590879	21921,4007057107	23422,1046728775
01.06.2010	21691,1459590879	21921,4007057107	23422,1046728775	25870,9257296058
01.07.2010	21921,4007057107	23422,1046728775	25870,9257296058	28649,7434237298
01.08.2010	23422,1046728775	25870,9257296058	28649,7434237298	31005,4236204159
01.09.2010	25870,9257296058	28649,7434237298	31005,4236204159	32261,7731788975
01.10.2010	28649,7434237298	31005,4236204159	32261,7731788975	32021,173884331
01.11.2010	31005,4236204159	32261,7731788975	32021,173884331	30294,0217076448
01.12.2010	32261,7731788975	32021,173884331	30294,0217076448	27513,1139417255
01.01.2011	32021,173884331	30294,0217076448	27513,1139417255	24423,4715263334
01.02.2011	30294,0217076448	27513,1139417255	24423,4715263334	21875,8798314448
01.03.2011	27513,1139417255	24423,4715263334	21875,8798314448	20582,9624529107
01.04.2011	24423,4715263334	21875,8798314448	20582,9624529107	20910,0045952944
01.05.2011	21875,8798314448	20582,9624529107	20910,0045952944	22764,2288484679
01.06.2011	20582,9624529107	20910,0045952944	22764,2288484679	25617,7875452123
01.07.2011	20910,0045952944	22764,2288484679	25617,7875452123	28659,7137325685
01.08.2011	22764,2288484679	25617,7875452123	28659,7137325685	31032,6809137336
01.09.2011	25617,7875452123	28659,7137325685	31032,6809137336	32083,9619308408
01.10.2011	28659,7137325685	31032,6809137336	32083,9619308408	31554,9353180312
01.11.2011	31032,6809137336	32083,9619308408	31554,9353180312	29651,7626259188
01.12.2011	32083,9619308408	31554,9353180312	29651,7626259188	26976,0443268622
01.01.2012	31554,9353180312	29651,7626259188	26976,0443268622	24337,2368955682
01.02.2012	29651,7626259188	26976,0443268622	24337,2368955682	

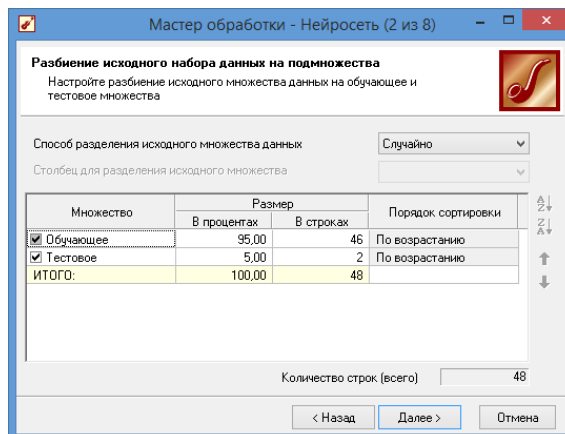
Теперь в качестве входных факторов можно использовать "Потребление эл. энергии - 12", "Потребление эл. энергии - 11" - данные по количеству потребляемой энергии 12 и 11 месяцев назад (относительно прогнозируемого месяца), и остальные. В качестве выходного поля можно использовать столбец "Потребление эл. энергии".

### Обучение нейросети

Построение модели прогноза.

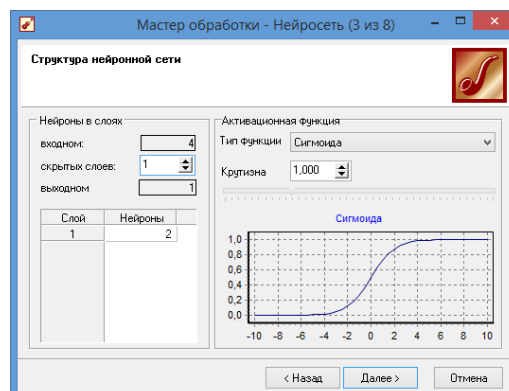
Открыть Мастер обработки и выбрать в нем нейронную сеть. На втором шаге Мастера необходимо установить входные поля. Например можно взять "Потребление эл. энергии - 4", "Потребление эл. энергии - 3" "Потребление эл. энергии - 2" и "Потребление эл. энергии - 1" а в качестве выходного - "Потребление эл. энергии". Остальные поля сделать информационными.

На следующем шаге необходимо указать разбиение тестового и обучающего множеств.



Множество	Размер		Порядок сортировки
	В процентах	В строках	
<input checked="" type="checkbox"/> Обучающее	95,00	46	По возрастанию
<input checked="" type="checkbox"/> Тестовое	5,00	2	По возрастанию
ИТОГО:	100,00	48	

На следующем этапе отмечаются необходимое количество слоев и нейронов в нейросети.



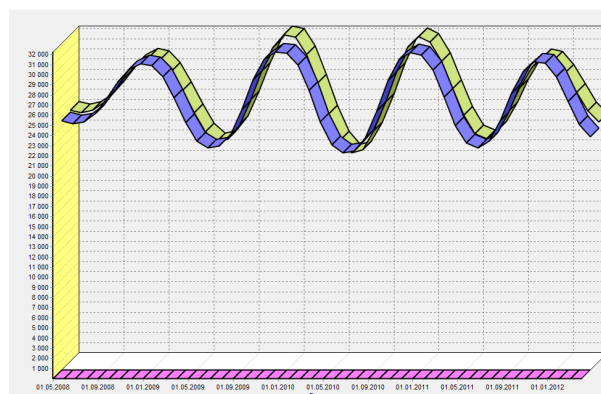
Слой	Нейроны
1	2

Перейдя далее, выбирается алгоритм обучения нейросети.

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

В Мастере настройки диаграммы выбрать для отображения поля " Потребление эл. энергии " и " Потребление эл. энергии \_OUT" - реальное и спрогнозированное значение.

Результатом будет два графика.

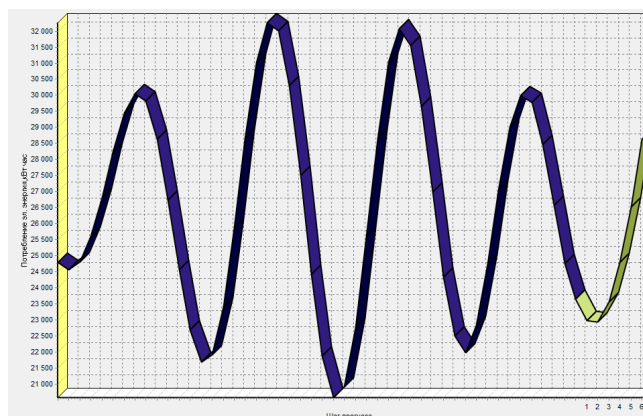


## Построение прогноза

Нейросеть обучена, осталось получить требуемый прогноз. Для этого необходимо открыть Мастер обработки и выбирать появившийся теперь обработчик "Прогнозирование".

На втором шаге Мастера предлагается настроить связи столбцов для прогнозирования временного ряда: откуда брать данные для столбца при очередном шаге прогноза.

Мастер сам верно настроил все переходы, поэтому остается только указать горизонт прогноза (на сколько вперед будем прогнозировать), а также для наглядности следует добавить к прогнозу исходные данные, установив в Мастере соответствующий флажок.



## Результат

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

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

Теперь аналитик может узнать количество потребленной электроэнергии в следующем месяце и шесть месяцев спустя.

### 2.3 Решение задачи прогнозирования временного ряда в Deductor

Подготовить данные для прогнозирования временного ряда в формате \*.txt

Провести анализ временного ряда

Провести прогнозирование временного ряда на 12 месяцев.

Сделать выводы

## Приложение А

Пример данных для прогнозирования временного ряда

Дата	Потребление эл. энергия,кВт час
01.01.2008	25668
01.02.2008	23292
01.03.2008	25155
01.04.2008	24228
01.05.2008	21510
01.06.2008	18513
01.07.2008	20960,95
01.08.2008	20079
01.09.2008	21951
01.10.2008	31212
01.11.2008	32688
01.12.2008	30438
01.01.2009	28764
01.02.2009	27387
01.03.2009	27171
01.04.2009	23037,669
01.05.2009	20960,95
01.06.2009	20361,525
01.07.2009	21972,3
01.08.2009	20659,725
01.09.2009	22602,975
01.10.2009	26337,15
01.11.2009	26279,25
01.12.2009	36532,425
01.01.2010	29880,45
01.02.2010	32113,125
01.03.2010	30285,675
01.04.2010	23484,975
01.05.2010	185127,5
01.06.2010	18588,675
01.07.2010	23471,475
01.08.2010	24175,5
01.09.2010	21903,825
01.10.2010	28014,15
01.11.2010	33909
01.12.2010	32032,35
01.01.2011	30678,75
01.02.2011	27801
01.03.2011	28182,225
01.04.2011	26865,375
01.05.2011	26390,55



01.06.2011	15997,5
01.07.2011	23166,45
01.08.2011	20960,95
01.09.2011	27836,175
01.10.2011	32928,75
01.11.2011	31778,85
01.12.2011	32908,425
01.01.2012	31761,3
01.02.2012	31200,45
01.03.2012	29871,3
01.04.2012	25868,4