

Лабораторная работа №4

СЕГМЕНТАЦИЯ КЛИЕНТОВ ТЕЛЕКОММУНИКАЦИОННОЙ КОМПАНИИ

Описание бизнес-задачи

В такой высокотехнологичной отрасли, как телекоммуникации, методы и подходы Data Mining получили широкое применение. Решаемые задачи прежде всего связаны с программами лояльности и удержанием существующей клиентской базы, а также с привлечением новых потребителей услуг.

В биллинговых системах телекоммуникационных компаний накапливаются большие объемы данных. В первую очередь это информация об абонентах и статистика использованных услуг. Анализ такой информации ручными и полуручными методами малоэффективен.

Постановка задачи. Руководство филиала региональной телекоммуникационной компании, предоставляющей услуги мобильной связи, поставило задачу сегментации абонентской базы. Ее целями являются:

- ☐ построение профилей абонентов путем выявления их схожего поведения в плане частоты, длительности и времени звонков, а также ежемесячных расходов;
- ☐ оценка наиболее и наименее доходных сегментов.

Эта информация может в дальнейшем использоваться для:

- ☐ разработки маркетинговых акций, направленных на определенные группы клиентов;
- ☐ разработки новых тарифных планов;
- ☐ оптимизации расходов на адресную SMS-рассылку о новых услугах и тарифах;
- ☐ предотвращения оттока клиентов в другие компании.

Данные за последние несколько месяцев, взятые из биллинговой системы, представляют собой таблицу со следующими полями (табл. 16.1).

Данные за последние несколько месяцев, взятые из биллинговой системы, представляют собой таблицу со следующими полями (табл. 16.1).

Таблица 16.1. Данные по абонентам из биллинговой системы

№	Поле	Описание	Тип
1	Возраст	Возраст клиента	Целый
2	Среднемесячный расход	Сколько в среднем денег в месяц тратит абонент на мобильную связь	Вещественный
3	Средняя продолжительность разговора	Сколько в среднем минут на исходящие звонки тратит абонент за месяц	Вещественный
4	Звонков днем за месяц	Количество исходящих звонков в утреннее и дневное время	Целый
5	Звонков вечером за месяц	Количество исходящих звонков в вечернее время	Целый
6	Звонков ночью за месяц	Количество исходящих звонков в ночное время	Целый
7	Звонки в другие города	Количество исходящих звонков в другие города	Целый
8	Звонки в другие страны	Число исходящих международных звонков	Целый
9	Доля звонков на стационарные телефоны	—	Вещественный
10	Количество SMS	Число исходящих SMS-сообщений в месяц	Целый

Исходные данные. Были отобраны только активные абоненты, которые регулярно пользовались услугами сотовой связи в течение последних нескольких месяцев. Данные находятся на прилагаемом к книге компакт-диске в файле `mobile.txt`.

Решение задачи

Покажем последовательность решения бизнес-задачи сегментации абонентов с помощью подхода, который основан на алгоритме Кохонена. Решение состоит из двух шагов:

- ☐ кластеризации объектов алгоритмом Кохонена;
- ☐ построения и интерпретации карты Кохонена.

В программе Deductor сети и карты Кохонена реализованы в обработчике **Карта Кохонена**, где содержатся сам алгоритм Кохонена и специальный визуализатор **Карта Кохонена**.

В Deductor канонический алгоритм Кохонена дополнен рядом возможностей.

- ❑ Алгоритм Кохонена применяется к сети Кохонена, состоящей из ячеек, упорядоченных на плоскости. По умолчанию размер карты равен 16×12 , что соответствует 192 ячейкам. В выходном наборе данных алгоритм Кохонена формирует поля Номер ячейки и Расстояние до центра ячейки.
- ❑ Ячейки карты с помощью специальной дополнительной процедуры объединяются в кластеры. Эта процедура — алгоритмы k-means и G-means. В результате в выходном наборе данных формируются поля Номер кластера и Расстояние до центра кластера.
- ❑ Каждый входной признак может иметь весовой коэффициент от 0 до 100 %, который влияет на расчет евклидова расстояния между векторами.

Импортируйте в Deductor набор данных из файла `mobile.txt`. Запустите Мастер обработки и выберите узел Карта Кохонена. Установите все поля, кроме Код, входными (рис. 16.1).

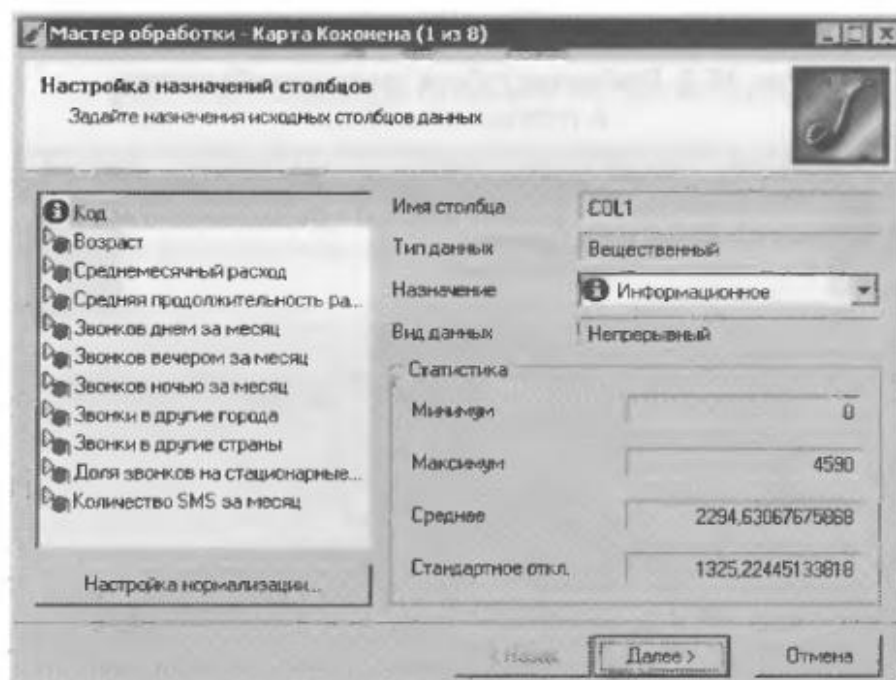


Рис. 16.1. Установка входных полей в алгоритме Кохонена

На этой же вкладке при нажатии кнопки **Настройка нормализации** откроется окно, где можно задать значимость каждого входного поля. Оставьте значимость всех полей без изменений.

Поскольку любой метод кластеризации, в том числе алгоритм Кохонена, субъективен, смысл в выделении отдельного тестового множества, как правило, отсутствует. Оставьте в обучающем множестве 100 % записей (рис. 16.2).

На третьей вкладке задаются размер и форма карты Кохонена (рис. 16.3). Увеличьте размер карты до 24×18 (соотношение рекомендуется делать кратным 4 : 3).

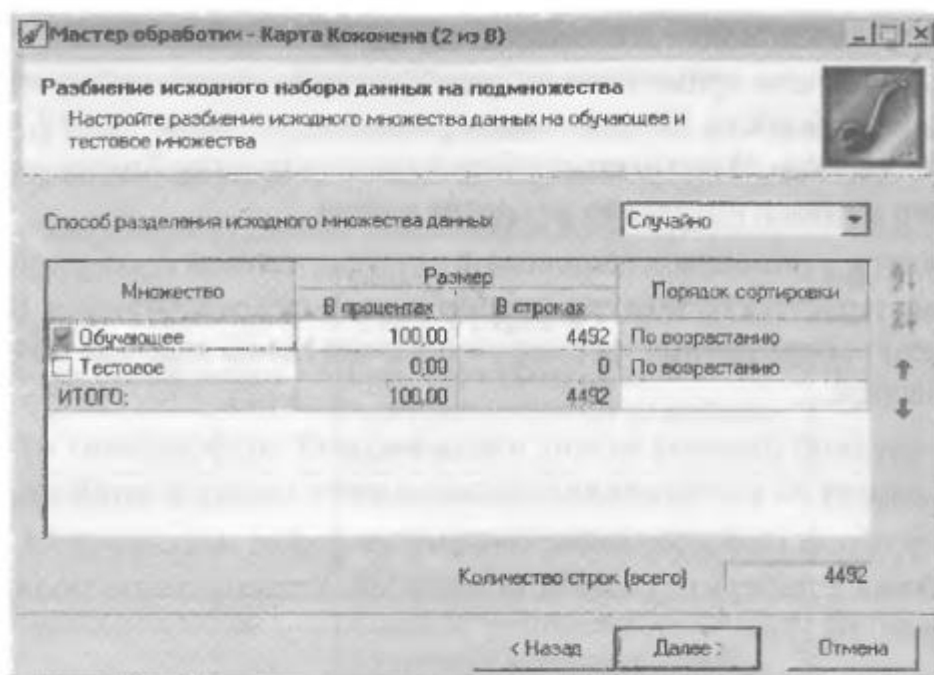


Рис. 16.2. Разбиение набора данных на обучающее и тестовое множества

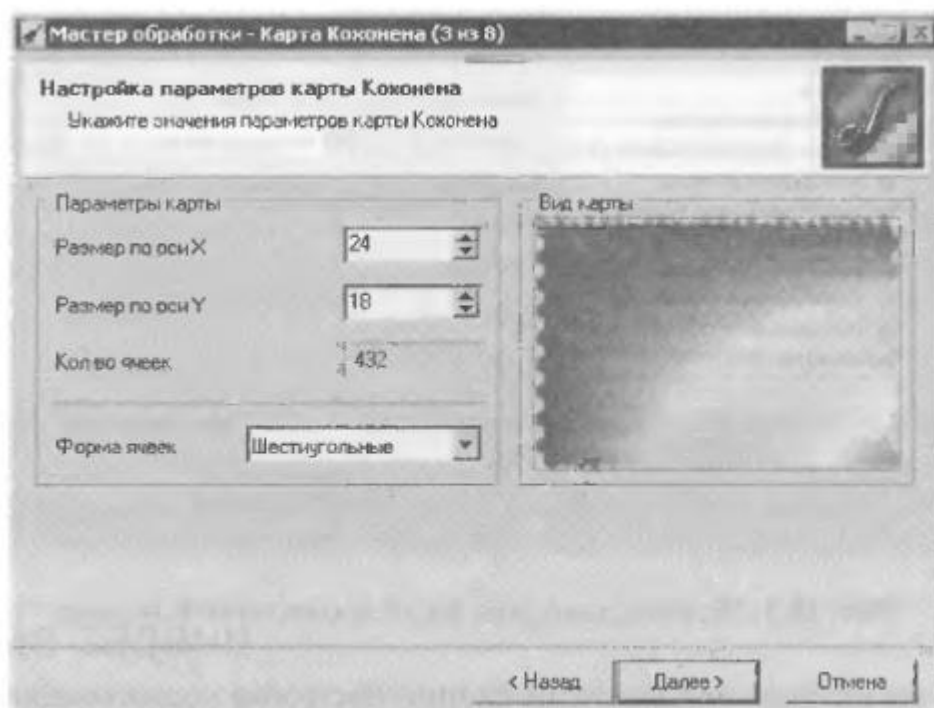


Рис. 16.3. Параметры будущей карты Кохонена

На следующем шаге оставьте все без изменений (рис. 16.4).

Наконец, на последнем шаге, предшествующем обучению, настраиваются параметры обучения алгоритма Кохонена (рис. 16.5).

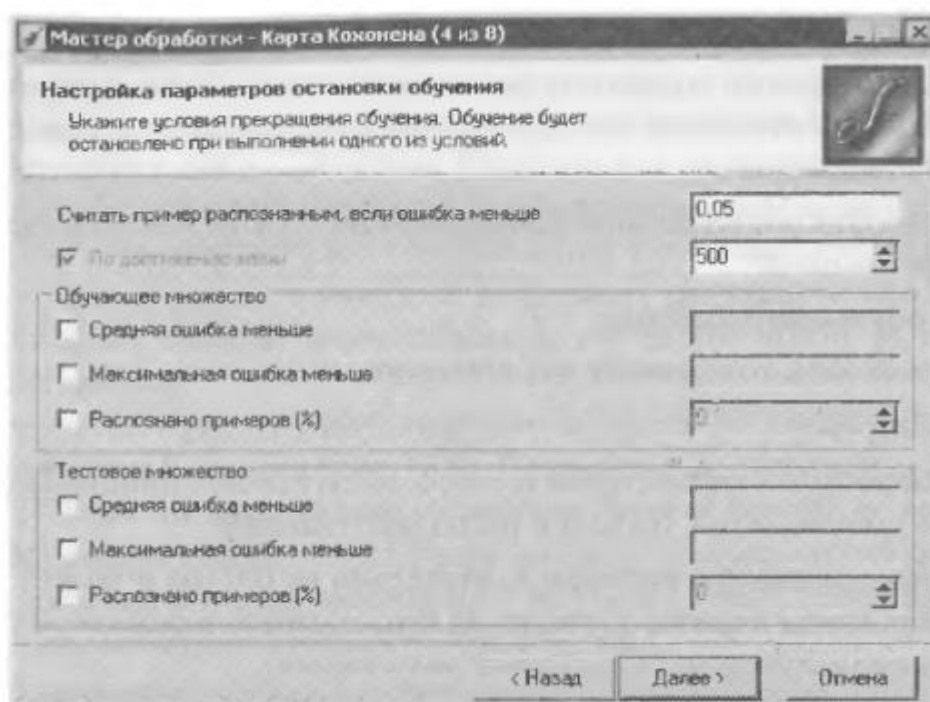


Рис. 16.4. Параметры остановки алгоритма Кохонена

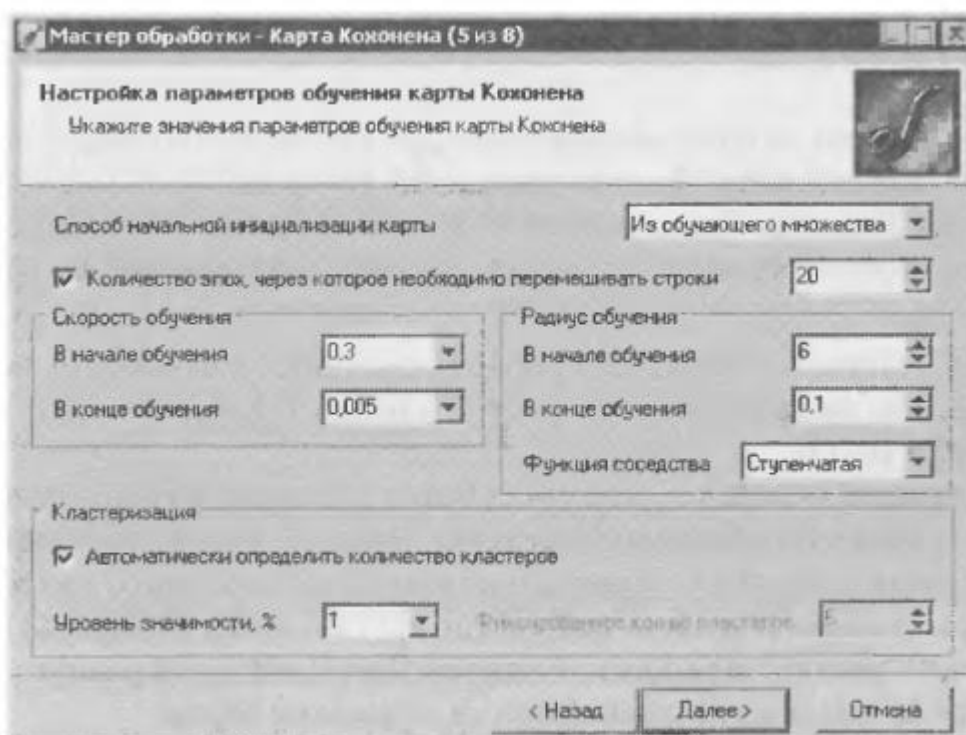


Рис. 16.5. Параметры обучения сети Кохонена

Здесь задаются следующие опции.

Способ начальной инициализации карты определяет, как будут установлены начальные веса нейронов карты. Удачно выбранный способ инициализации может

существенно ускорить обучение и привести к получению более качественных результатов. Доступны три варианта.

- ❑ **Случайными значениями** — начальные веса нейронов будут инициализированы случайными значениями.
- ❑ **Из обучающего множества** — в качестве начальных весов будут использоваться случайные примеры из обучающего множества.
- ❑ **Из собственных векторов** — начальные веса нейронов карты будут проинициализированы значениями подмножества гиперплоскости, через которую проходят два главных собственных вектора матрицы ковариации входных значений обучающей выборки.

При выборе способа начальной инициализации нужно руководствоваться следующей информацией:

- ❑ объемом обучающей выборки;
- ❑ количеством эпох, отведенных для обучения;
- ❑ размером карты.

Между указанными параметрами и способом начальной инициализации существует много зависимостей. Выделим несколько главных.

- ❑ Если объем обучающей выборки значительно (в 100 раз и более) превышает число ячеек карты и время обучения не играет первоочередной роли, то лучше выбрать инициализацию случайными значениями.
- ❑ Если объем обучающей выборки не очень велик, время обучения ограничено или если необходимо уменьшить вероятность появления после обучения пустых ячеек, в которые не попало ни одного экземпляра обучающей выборки, то следует использовать инициализацию примерами из обучающего множества.
- ❑ Инициализацию из собственных векторов можно использовать при любом стечении обстоятельств. Именно этот способ лучше выбирать при первом ознакомлении с данными. Единственное замечание: вероятность появления пустых ячеек после обучения выше, чем при инициализации примерами из обучающего множества.

Скорость обучения — задается скорость обучения в начале и в конце обучения сети Кохонена. Рекомендуемые значения: 0,1–0,3 в начале обучения и 0,05–0,005 в конце.

Радиус обучения — задается радиус обучения в начале и в конце обучения сети Кохонена, а также тип функции соседства. Вначале радиус обучения должен быть достаточно большим — примерно половина размера карты (максимальное линейное расстояние от любого нейрона до другого любого нейрона) или меньше, а в конце — достаточно малым, 1 или меньше. Начальный радиус в Deductor подбирается автоматически в зависимости от размера карты.

В этом же блоке задается вид функции соседства: гауссова или ступенчатая. Если функция соседства ступенчатая, то «соседями» нейрона-победителя будут считаться все нейроны, линейное расстояние до которых не больше текущего радиуса обучения. Если применяется гауссова функция соседства, то «соседями» нейрона-победителя будут считаться все нейроны карты, но в разной степени. При использовании гауссовой функции соседства обучение проходит более плавно и равномерно, так как одновременно изменяются веса всех нейронов, что может дать немного лучший результат, чем если бы использовалась ступенчатая функция. Однако и времени на обучение требуется больше, поскольку в каждой эпохе корректируются все нейроны.

Кластеризация — в этой области указываются параметры алгоритма k-means, который запускается после алгоритма Кохонена для кластеризации ячеек карты. Здесь нужно либо позволить алгоритму автоматически определить число кластеров, либо сразу зафиксировать его. Следует знать, что автоматически подбираемое число кластеров не всегда приводит к желаемому результату: оно может быть слишком большим, поэтому рассчитывать на эту опцию можно только на этапе исследования данных.

Нажмите кнопку **Пуск** — в следующем окне можно будет увидеть динамику процесса обучения сети Кохонена (рис. 16.6). По умолчанию алгоритм делает 500 итераций (эпох). Если предварительно установить флажок **Рестарт**, то веса нейронов будут проинициализированы согласно выбранному на предыдущем шаге способу инициализации, иначе обучение начнется с текущих весовых коэффициентов (это справедливо только при повторной настройке узла).

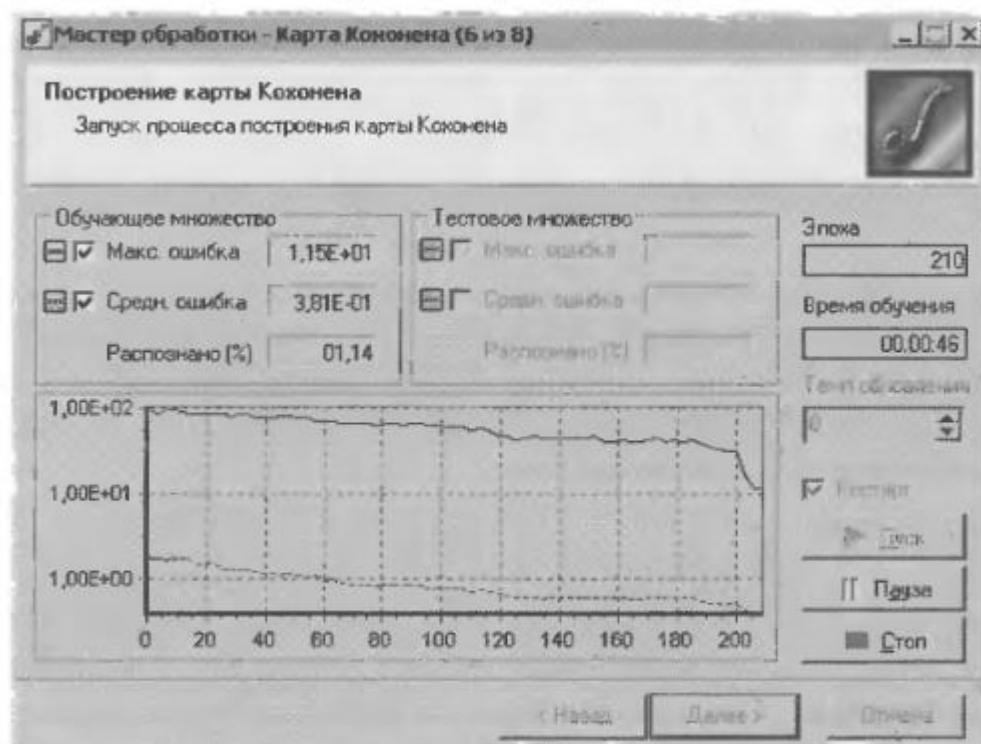


Рис. 16.6. Обучение сети Кохонена

Для обученной сети Кохонена предлагается специализированный визуализатор — **Карта Кохонена**. Параметры ее отображения задаются на одноименной вкладке мастера (рис. 16.7).

Область **Список допустимых отображений карты** содержит три группы — **Входные столбцы**, **Выходные столбцы** и **Специальные**. Последние не связаны с каким-либо полем набора данных, а служат для анализа всей карты.

- ☐ **Матрица расстояний** применяется для визуализации структуры кластеров, полученных в результате обучения карты. Большое значение говорит о том, что данный нейрон сильно отличается от окружающих и относится к другому классу.
- ☐ **Матрица ошибок квантования** отображает среднее расстояние от расположения примеров до центра ячейки. Расстояние считается как евклидово. Матрица ошибок квантования показывает, насколько хорошо обучена сеть Кохонена. Чем меньше среднее расстояние до центра ячейки, тем ближе к ней расположены примеры и тем лучше модель.
- ☐ **Матрица плотности попадания** отображает количество объектов, попавших в ячейку.
- ☐ **Кластеры** — ячейки карты Кохонена, объединенные в кластеры алгоритмом k-means.

- ❑ **Проекция Саммона** — матрица, являющаяся результатом проецирования многомерных данных на плоскость. При этом данные, расположенные рядом в исходной многомерной выборке, будут расположены рядом и на плоскости.

Дополнительно справа имеется еще ряд настроек.

- ❑ **Способ раскрашивания ячеек** — цветная палитра или градация серого. Цветная палитра нагляднее, однако если потребуется встраивать карту Кохонена в отчет с последующей распечаткой на бумажном носителе, то лучше выбрать серую цветовую схему.
- ❑ **Сглаживание цветов карты** — цвета на картах будут сглажены, то есть будет обеспечен более плавный переход цветов. Это поможет устранить случайные выбросы.

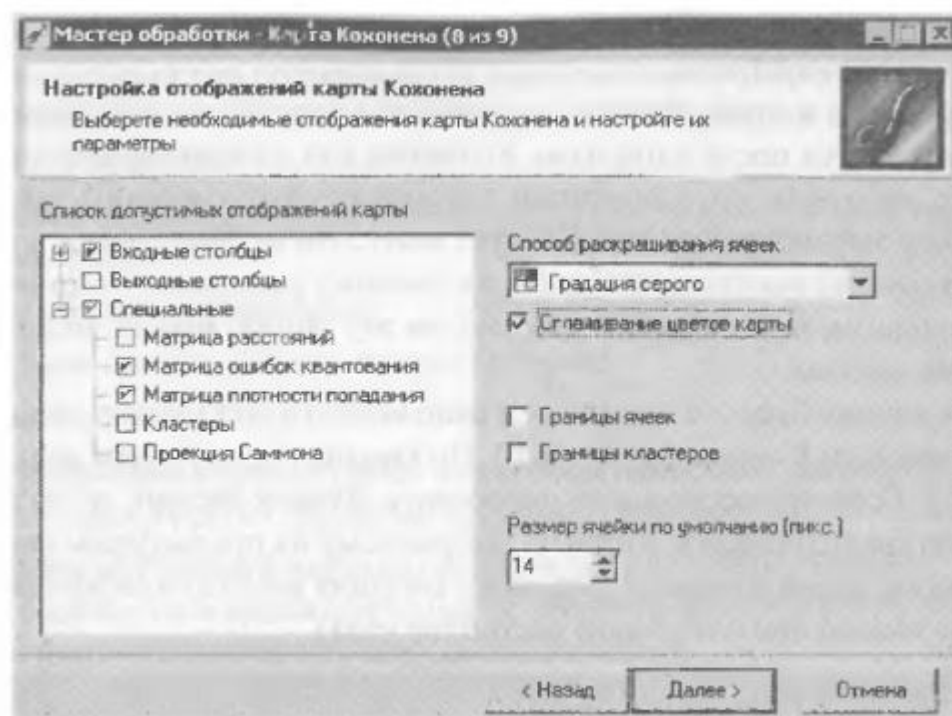


Рис. 16.7. Параметры карты Кохонена

- ☐ Границы ячеек — установка данного флажка позволит включить отображение границ ячеек на карте.
- ☐ Границы кластеров — установка данного флажка позволит включить отображение границ кластеров на всех картах. Этот режим удобен для анализа структуры кластеров.
- ☐ Размер ячейки по умолчанию — указывается размер ячейки на карте в пикселах (по умолчанию 16).

Текущая ячейка отображается на карте маленькой окружностью черного цвета. Изменить текущую ячейку просто: щелкнуть кнопкой мыши на нужном участке карты. Внизу каждой карты на градиентной шкале в желтом прямоугольнике отображается числовое значение признака, соответствующее цвету ячейки.

На рис. 16.8 приведены получившиеся карты Кохонена. По матрице плотности попадания видно, что в одной ячейке сосредоточилось 259 объектов. Эта ячейка выделяется белым цветом. Можно приступить к интерпретации результатов кластеризации.

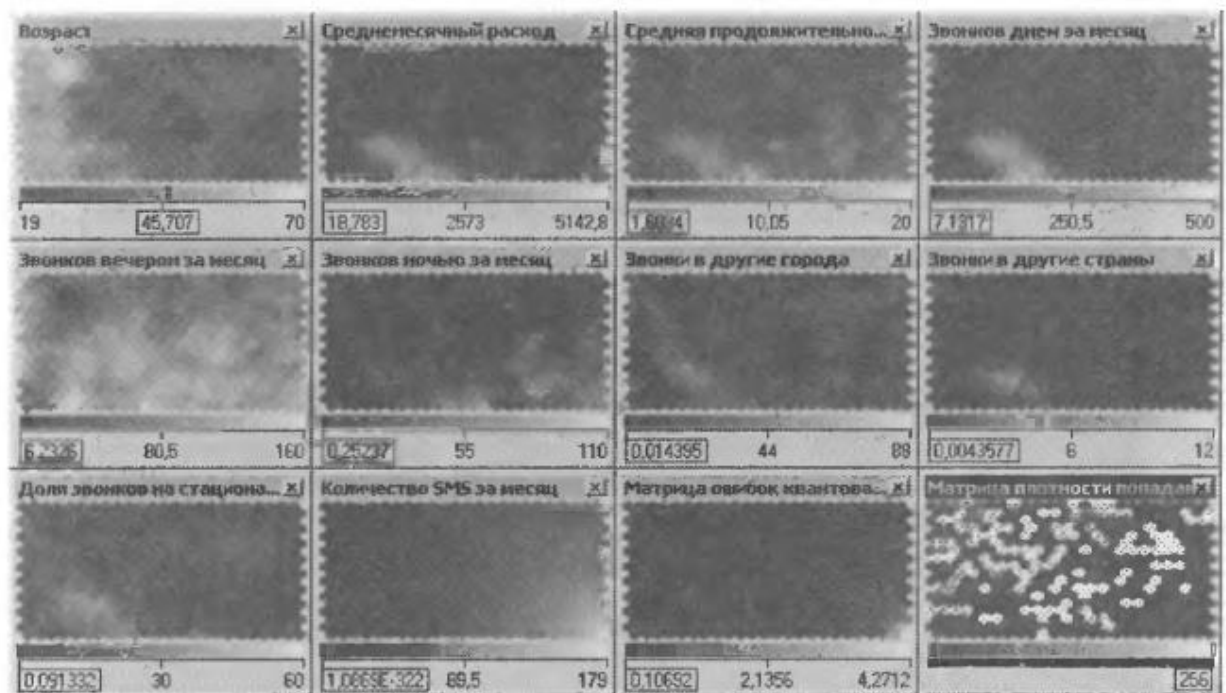


Рис. 16.8. Карты Кохонена для сегментации абонентов сети сотовой связи

Попробуем выделить на карте изолированные области самостоятельно (без использования встроенного метода группировки ячеек алгоритмом k-means).

Анализируя карту **Возраст** (рис. 16.9), видим, что четко выделяются три возрастные группы: молодежь, люди среднего возраста и люди старше 45 лет.

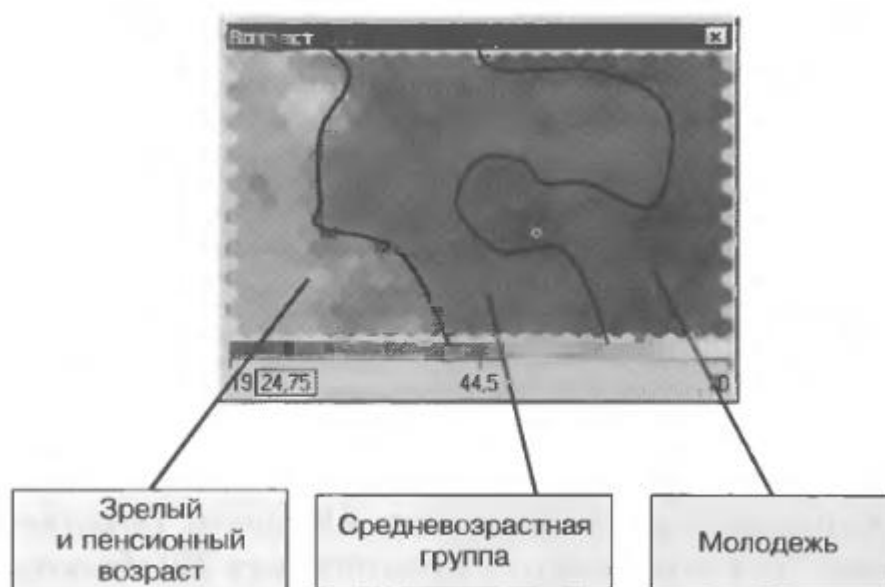


Рис. 16.9. Деление по возрастным группам

Остановимся подробнее на молодежи. Она неоднородна — здесь можно выделить несколько кластеров. Первый расположен в правом нижнем углу (рис. 16.10). Абоненты этой условной зоны активно и продолжительно разговаривают по телефону вечером и ночью, отправляют много SMS-сообщений, соответственно, и тратят на разговоры больше денег, чем другие представители возрастной группы. Обратите внимание, что в этот кластер попала львиная доля тех, кто увлекается ночными разговорами. Можно предположить, что это студенты и молодежь, часто проводящие вечера вне дома.

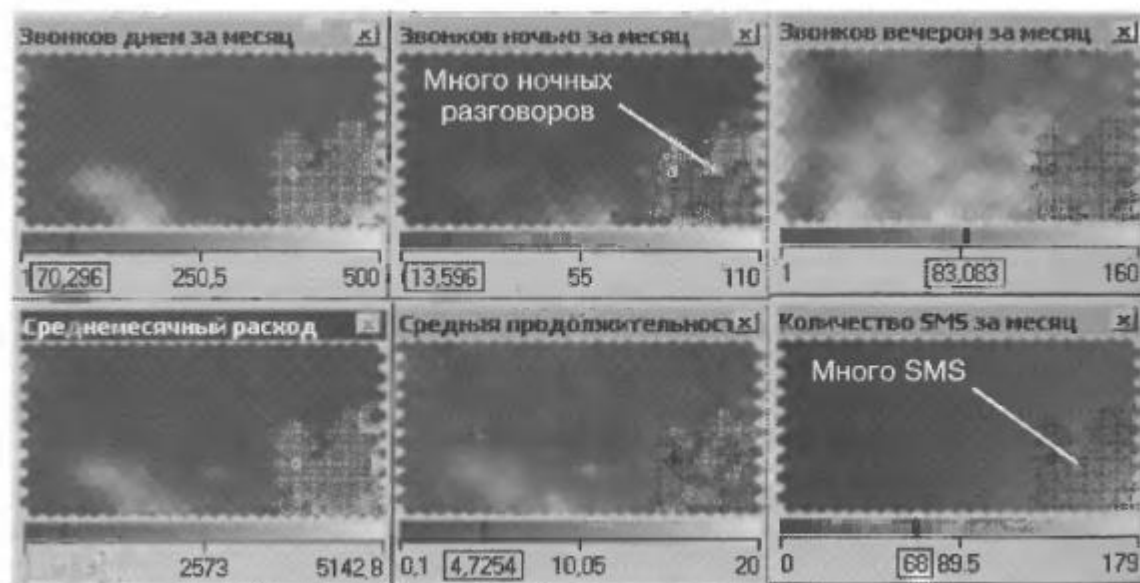


Рис. 16.10. Кластер «Активная молодежь»

Вверху (рис. 16.11) сосредоточилась небольшая по числу ячеек группа молодежи, которая не отличается активностью разговоров и пользования SMS-услугами ни днем, ни вечером, ни тем более ночью, и, как следствие, ежемесячные расходы на связь у представителей этого кластера невелики.

Остальные люди в этой возрастной группе ничем особенным не выделяются: умеренные расходы на связь и преимущественно вечерние разговоры. Можно предположить, что сюда попала наибольшая часть молодежи. Таким образом, в молодежной возрастной группе мы обнаружили три кластера.

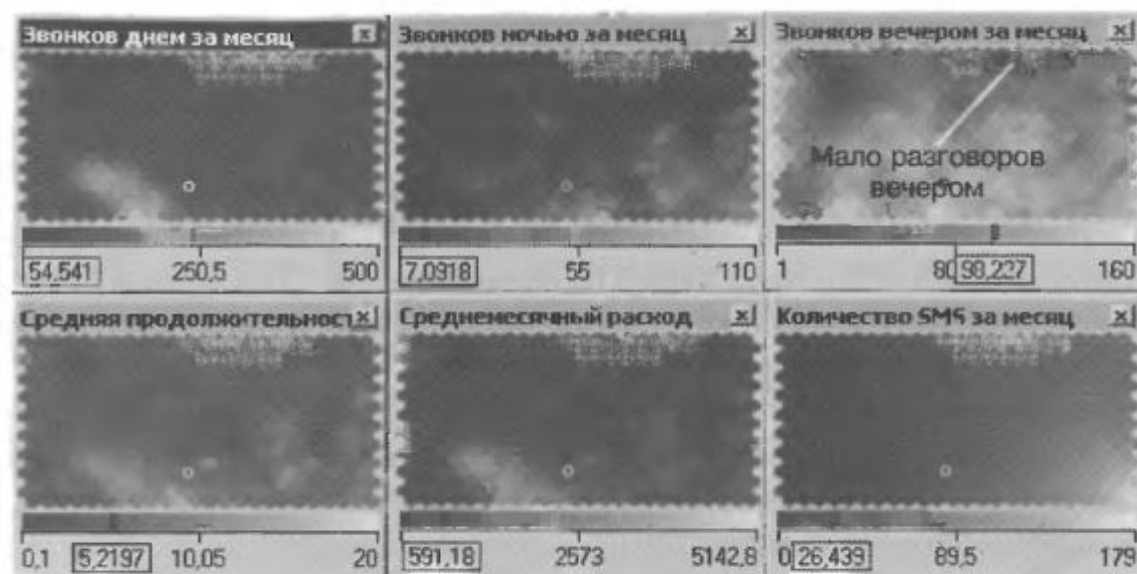


Рис. 16.11. Кластер молодежи с пониженным потреблением услуг связи

Продолжим интерпретацию карты Кохонена и теперь остановимся на людях зрелого и пенсионного возраста. Обратим внимание на ярко выраженный сгусток в нижней области, в котором практически по всем признакам, кроме SMS, наблюдаются высокие значения, в том числе по звонкам в другие города и страны (рис. 16.12). Это так называемые VIP-клиенты: бизнесмены, руководители, топ-менеджеры. Они преимущественно зрелого возраста, очень много разговаривают днем и вечером (скорее всего, по работе) и практически не пользуются SMS-услугами. Месячные расходы на связь у этой категории абонентов самые высокие.

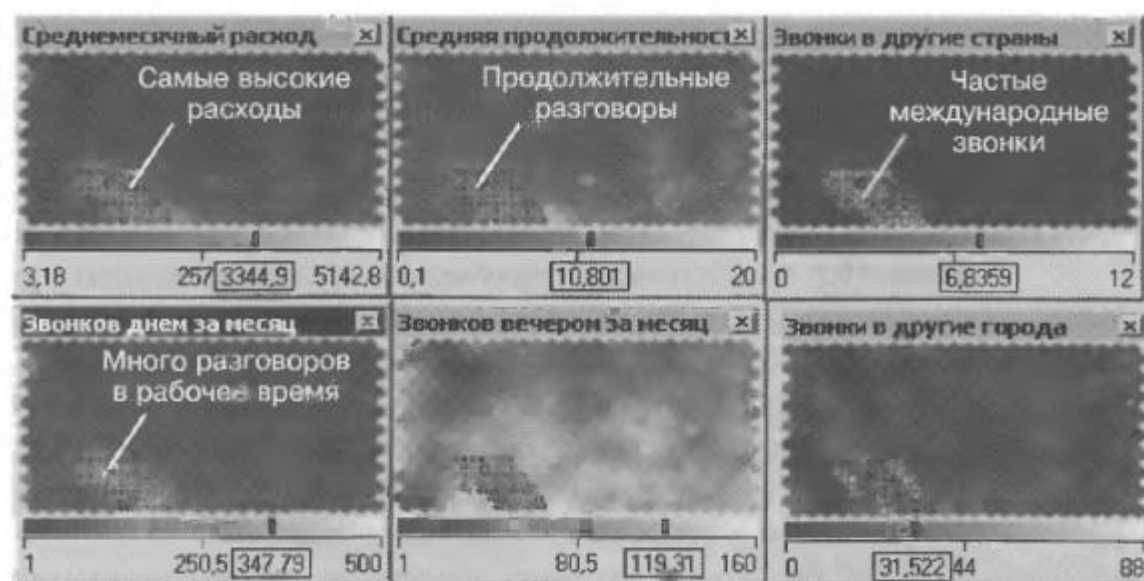




Рис. 16.12. Кластер «VIP-клиенты»

Чуть выше в небольшом кластере наблюдается противоположная картина: люди практически не пользуются услугами сотовой сети (рис. 16.13). Вероятнее всего, это пенсионеры, которым мобильная связь нужна преимущественно для приема входящих звонков, сами же они почти не звонят. Их расходы на связь самые низкие, возможно, из-за того, что единственным их доходом является пенсия.

Изучим статистические характеристики этой группы людей.

Нажмите кнопку Показать/скрыть окно данных , затем — кнопку Изменить способ фильтрации и установите Фильтр по выделенному . Там же переключитесь в режим статистики (кнопка Способ отображения). Колонка Среднее даст следующие вычисленные значения (табл. 16.2).

Остальных людей в возрастной группе «Зрелый и пенсионный возраст» объединяет то, что они в основном звонят вечером и не используют SMS-сервис. С большой долей вероятности можно утверждать, что сюда входят работающие пенсионеры, дачники, родители совершеннолетних детей.



Рис. 16.13. Пенсионеры, практически не делающие исходящих звонков

Таблица 16.2. Статистические характеристики кластера 5

№	Признак	Среднее значение
1	Возраст	64,5
2	Среднемесячный расход	49,7
3	Средняя продолжительность разговоров, мин	2,1
4	Звонков днем за месяц	13,4
5	Звонков вечером за месяц	7,4
6	Звонков ночью за месяц	0,3
7	Звонки в другие города	0,5

№	Признак	Среднее значение
8	Звонки в другие страны	0,05
9	Доля звонков на стационарные телефоны, %	5,7
10	Количество SMS в месяц	1,6

Осталась последняя, средневозрастная группа. Это кластер работающих людей. В нем можно отметить группу тех, кто совершает мало звонков вечером.

Теперь включите автоматическую группировку ячеек в кластеры: Настроить отображения — Кластеры. При установленном флажке Автоматически определить количество кластеров (с уровнем значимости 1,00) будет работать алгоритм G-means и получится 11 кластеров (рис. 16.14, а). Это очень много, поэтому следует принудительно установить, скажем, 6 кластеров (рис. 16.14, б).

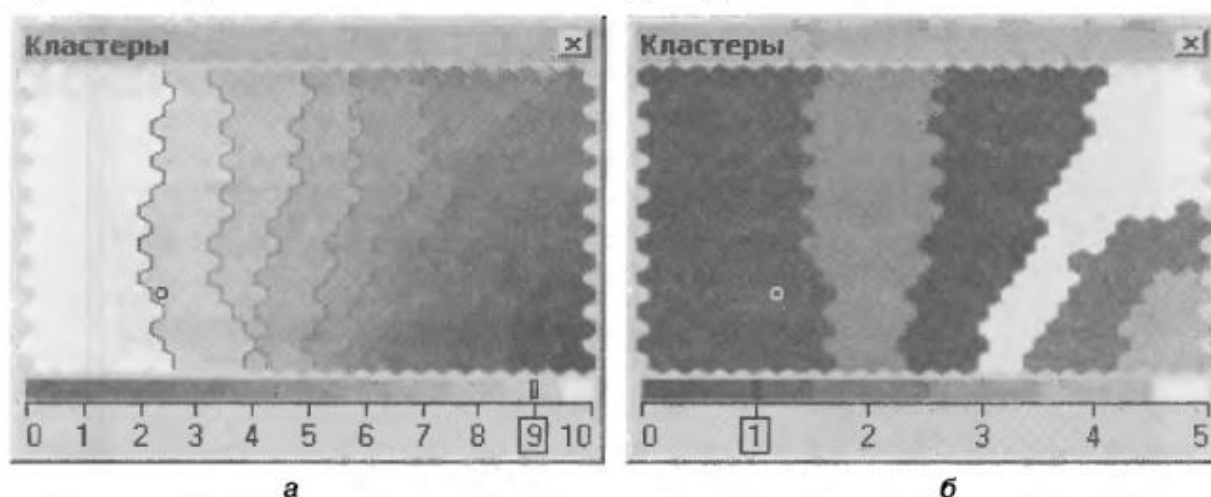


Рис. 16.14. Варианты группировки ячеек:
a — алгоритм G-means; *б* — алгоритм k-means

Видно, что алгоритм k-means при 6 кластерах выделил целиком кластер «Зрелый и пенсионный возраст», раздробились группы «Молодежь» и «Люди среднего возраста». Не были явно выделены кластеры 2, 4 и 5 из табл. 16.3. Тем не менее автоматическая группировка ячеек в Deductor имеет одно важное преимущество: в наборе данных появляется столбец Кластер с номером, который можно использовать в дальнейшем, в частности «прогонять» новые объекты и получать для них номер кластера.

Кроме того, через визуализатор Профили кластеров можно изучить статистические характеристики кластеров: мощность (число записей, попавших в кластер), среднее и др. (рис. 16.15). Это очень полезно, так как карты Кохонена не позволяют увидеть все детали. Кластерам можно дать краткие названия.

В этом визуализаторе особый интерес представляют две характеристики — значимость и доверительный интервал. Значимость выражается в процентах, и для непрерывных полей используется *t*-критерий Стьюдента. 95%-ный доверительный интервал представляется в виде графического изображения для среднего значения кластера.

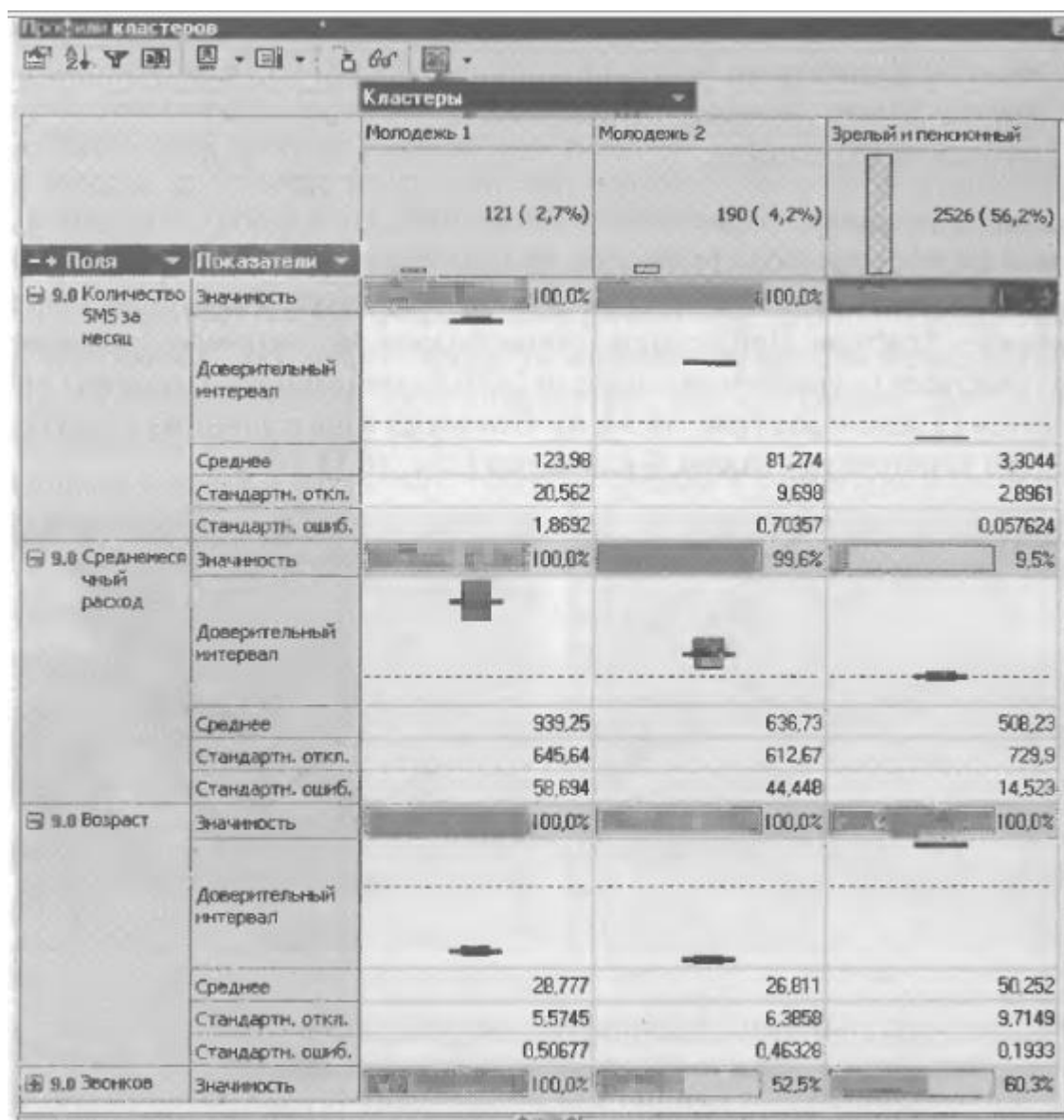


Рис. 16.15. Профили кластеров

ЗАДАНИЕ.

Реализовать рассмотренный проект (см. постановку задачи). В отчет по лабораторной работе включить принт-скрины хода выполнения работы, сформировать и предоставить файлы проекта.