# ЛЕКЦИЯ 6. ЗАДАЧИ DATA MINING. ПРОГНОЗИРОВАНИЕ И ВИЗУАЛИЗАЦИЯ

Мы продолжаем рассматривать наиболее распространенные и востребованные задачи *Data Mining*. В этой лекции мы подробно остановимся на задачах *прогнозирования* и *визуализации*.

**Задача прогнозирования**

Задачи *прогнозирования* решаются в самых разнообразных областях человеческой деятельности, таких как наука, экономика, производство и множество других сфер. *Прогнозирование* является важным элементом организации управления как отдельными хозяйствующими субъектами, так и экономики в целом.

Развитие методов *прогнозирования* непосредственно связано с развитием информационных технологий, в частности, с ростом объемов хранимых данных и усложнением методов и алгоритмов *прогнозирования*, реализованных в инструментах *Data Mining*.

Задача *прогнозирования*, пожалуй, может считаться одной из наиболее сложных задач *Data Mining*, она требует тщательного исследования исходного набора данных и методов, подходящих для анализа.

***Прогнозирование*** (от греческого Prognosis), в широком понимании этого слова, определяется как опережающее отражение будущего. Целью *прогнозирования* является предсказание будущих событий.

***Прогнозирование*** (*forecasting*) является одной из задач *Data Mining* и одновременно одним из ключевых моментов при принятии решений.

Прогностика (*prognostics*) - теория и практика *прогнозирования*.

*Прогнозирование* направлено на *определение* тенденций динамики конкретного объекта или события на основе ретроспективных данных, т.е. анализа его состояния в прошлом и настоящем. Таким образом, решение задачи *прогнозирования* требует некоторой *обучающей выборки* данных.

***Прогнозирование*** - установление функциональной зависимости между зависимыми и независимыми переменными.

*Прогнозирование* является распространенной и востребованной задачей во многих областях человеческой деятельности. В результате *прогнозирования* уменьшается риск принятия неверных, необоснованных или субъективных решений.

Примеры его задач: прогноз движения денежных средств, *прогнозирование* урожайности агрокультуры, *прогнозирование* финансовой устойчивости предприятия.

Типичной в сфере маркетинга является задача *прогнозирования* рынков (market *forecasting*). В результате решения данной задачи оцениваются перспективы развития конъюнктуры определенного рынка, изменения рыночных условий на будущие периоды, определяются тенденции рынка (структурные изменения, потребности покупателей, изменения цен).

Обычно в этой области решаются следующие практические задачи:

- прогноз продаж товаров (например, с целью определения нормы товарного запаса);
- *прогнозирование* продаж товаров, оказывающих влияние друг на друга;
- прогноз продаж в зависимости от внешних факторов.

Помимо экономической и финансовой сферы, задачи *прогнозирования* ставятся в самых разнообразных областях: медицине, фармакологии; популярным сейчас становится политическое *прогнозирование*.

В самых общих чертах решение задачи *прогнозирования* сводится к решению таких подзадач:

- выбор модели *прогнозирования* ;
- анализ адекватности и точности построенного прогноза.

**Сравнение задач прогнозирования и классификации**

В предыдущей лекции нами была рассмотрена задача классификации. *Прогнозирование* сходно с задачей классификации.

Многие методы *Data Mining* используются для решения задач классификации и *прогнозирования*. Это, например, линейная регрессия, нейронные сети, деревья решений (которые иногда так и называют - деревья *прогнозирования* и классификации).

Задачи классификации и *прогнозирования* имеют сходства и различия.

Так в чем же сходство задач *прогнозирования* и классификации? При решении обеих задач используется двухэтапный процесс построения модели на основе обучающего набора и ее использования для предсказания неизвестных значений зависимой переменной.

Различие задач классификации и *прогнозирования* состоит в том, что в первой задаче предсказывается *класс* зависимой переменной, а во второй - числовые значения зависимой переменной, пропущенные или неизвестные (относящиеся к будущему).

Возвращаясь к примеру о туристическом агентстве, рассмотренном в предыдущей лекции, мы можем сказать, что определения класса клиента является решением задачи классификации, а *прогнозирование* дохода, который принесет этот клиент в будущем году, будет решением задачи *прогнозирования*.

**Прогнозирование и временные ряды**

Основой для *прогнозирования* служит историческая *информация*, хранящаяся в базе данных в виде *временных рядов*.

Существует понятие *Data Mining временных рядов* (Time-*Series Data Mining*).

Подробно с этим понятием можно ознакомиться в.

На основе ретроспективной информации в виде *временных рядов* возможно решение различных задач *Data Mining*. На рис. 6.1 представлены результаты опроса относительно *Data Mining временных рядов*. Как видим, наибольший *процент* (23%) среди решаемых задач занимает *прогнозирование*. Далее идут классификация и *кластеризация* (по 14%), *сегментация* и выявление аномалий (по 9%), обнаружение правил (8%). На другие задачи приходится менее чем по 6%.

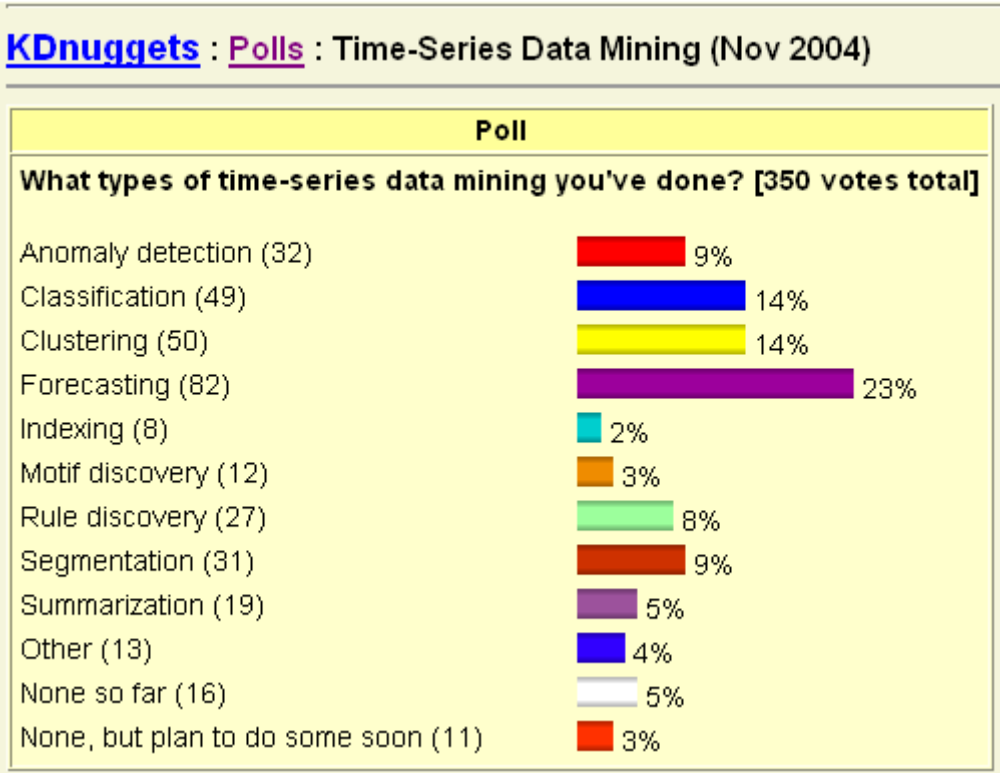


**Рис. 6.1.** Data Mining временных рядов

Однако чтобы сосредоточиться на понятии *прогнозирования*, мы будем рассматривать *временные ряды* лишь в рамках решения задачи *прогнозирования*.

Приведем два принципиальных отличия *временного ряда* от простой последовательности наблюдений:

- Члены *временного ряда*, в отличие от элементов *случайной выборки*, не являются статистически независимыми.
- Члены *временного ряда* не являются одинаково распределенными.

*Временной ряд* - последовательность наблюдаемых значений какого-либо признака, упорядоченных в неслучайные моменты времени.

Отличием анализа *временных рядов* от анализа случайных выборок является предположение о равных промежутках времени между наблюдениями и их хронологический порядок. Привязка наблюдений ко времени играет здесь ключевую роль, тогда как при анализе *случайной выборки* она не имеет никакого значения.

Типичный пример *временного ряда* - данные биржевых торгов.

*Информация*, накопленная в разнообразных базах данных предприятия, является *временными рядами*, если она расположена в хронологическом порядке и произведена в последовательные моменты времени.

*Анализ временного ряда* осуществляется с целью:

- определения природы ряда;
- *прогнозирования* будущих значений ряда.

В процессе определения структуры и закономерностей *временного ряда* предполагается обнаружение: шумов и выбросов, *тренда*, *сезонной компоненты*, *циклической компоненты*. *Определение* природы *временного*

*ряда* может быть использовано как своеобразная "разведка" данных. *Знание* аналитика о наличии *сезонной компоненты* необходимо, например, для определения количества записей выборки, которое должно принимать участие в построении прогноза.

*Шумы* и выбросы будут подробно обсуждаться в последующих лекциях курса. Они усложняют *анализ временного ряда*. Существуют различные методы определения и фильтрации выбросов, дающие возможность исключить их с целью более качественного *Data Mining*.

**Тренд, сезонность и цикл**

Основными составляющими *временного ряда* являются *тренд* и *сезонная компонента*. Составляющие этих рядов могут представлять собой либо *тренд*, либо *сезонную компоненту*.

*Тренд* является систематической компонентой *временного ряда*, которая может изменяться во времени.

***Трендом*** называют неслучайную функцию, которая формируется под действием общих или долговременных тенденций, влияющих на *временной ряд*.

Примером тенденции может выступать, например, фактор роста исследуемого рынка.

Автоматического способа обнаружения *трендов* во *временных рядах* не существует. Но если *временной ряд* включает монотонный *тренд* (т.е. отмечено его устойчивое возрастание или устойчивое убывание), анализировать *временной ряд* в большинстве случаев нетрудно.

Существует большое разнообразие постановок задач *прогнозирования*, которое можно подразделить на две группы: *прогнозирование* односерийных рядов и *прогнозирование* мультисерийных, или взаимовлияющих, рядов.

Группа прогнозирования односерийных рядов включает задачи построения прогноза одной переменной по ретроспективным данным только этой переменной, без учета влияния других переменных и факторов.

Группа *прогнозирования* мультисерийных, или взаимовлияющих, рядов включает задачи анализа, где необходимо учитывать взаимовлияющие факторы на одну или несколько переменных.

Кроме деления на классы по односерийности и многосерийности, ряды также бывают сезонными и несезонными.

Последнее деление подразумевает наличие или отсутствие у *временного ряда* такой составляющей как сезонность, т.е. включение *сезонной компоненты*.

Сезонная составляющая *временного ряда* является периодически повторяющейся компонентой *временного ряда*.

Свойство сезонности означает, что через примерно равные промежутки времени форма кривой, которая описывает поведение зависимой переменной, повторяет свои характерные очертания.

Свойство сезонности важно при определении количества ретроспективных данных, которые будут использоваться для *прогнозирования*.

Рассмотрим простой пример. На рис. 6.2. приведен фрагмент ряда, который иллюстрирует поведение переменной "объемы продажи товара X" за

период, составляющий один месяц. При изучении кривой, приведенной на рисунке, аналитик не может сделать предположений относительно повторяемости формы кривой через равные промежутки времени.

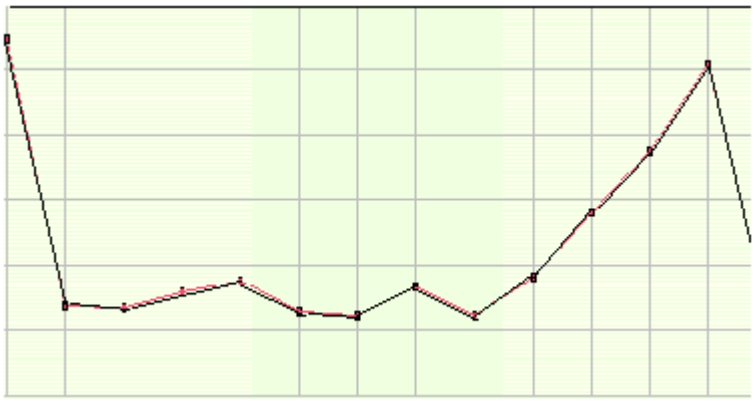

**Рис. 6.2.** Фрагмент временного ряда за сезонный период

Однако при рассмотрении более продолжительного ряда (за 12 месяцев), изображенного на рис. 6.3, можно увидеть явное наличие *сезонной компоненты*. Следовательно, о сезонности продаж можно говорить только, когда рассматриваются данные за несколько месяцев.

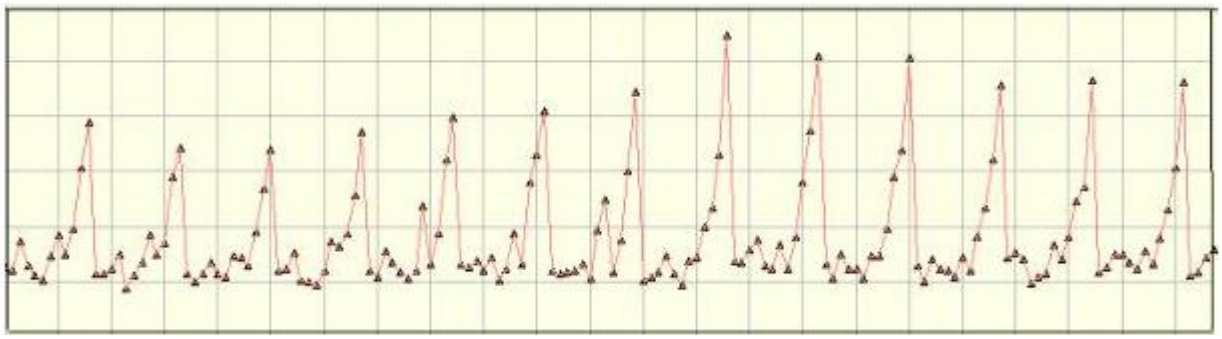

**Рис. 6.3.** Фрагмент временного ряда из 12-ти сезонных периодов

Таким образом, в процессе подготовки данных для *прогнозирования* аналитику следует определить, обладает ли ряд, который он анализирует, свойством сезонности.

Определение наличия компоненты сезонности необходимо для того, чтобы входная информация обладала свойством репрезентативности.

Ряд можно считать несезонным, если при рассмотрении его внешнего вида нельзя сделать предположений о повторяемости формы кривой через равные промежутки времени.

Иногда по внешнему виду кривой ряда нельзя определить, является он сезонным или нет.

Существует понятие сезонного мультиряда. В нем каждый ряд описывает поведение факторов, которые влияют на зависимую (целевую) переменную.

Пример такого ряда - ряды продаж нескольких товаров, подверженных сезонным колебаниям.

При сборе данных и выборе факторов для решения задачи по *прогнозированию* в таких случаях следует учитывать, что влияние объемов продаж товаров друг на друга здесь намного меньше, чем воздействие фактора сезонности.

Важно не путать понятия *сезонной компоненты* ряда и сезонов природы. Несмотря на близость их звучания, эти понятия разнятся. Так, например, объемы продаж мороженого летом намного больше, чем в другие сезоны, однако это является тенденцией спроса на данный товар.

Очень часто *тренд* и сезонность присутствуют во *временном ряде* одновременно.

Пример. Прибыль фирмы растет на протяжении нескольких лет (т.е. во *временном ряде* присутствует *тренд* ); ряд также содержит *сезонную компоненту*.

Отличия *циклической компоненты* от сезонной:

1. Продолжительность цикла, как правило, больше, чем один сезонный период;
2. Циклы, в отличие от сезонных периодов, не имеют определенной продолжительности.

При выполнении каких-либо преобразований понять природу *временного ряда* значительно проще, такими преобразованиями могут быть, например, удаление *тренда* и сглаживание ряда.

Перед началом *прогнозирования* необходимо ответить на следующие вопросы:

1. Что нужно прогнозировать?
2. В каких временных элементах (параметрах)?
3. С какой точностью прогноза?

При ответе на первый вопрос, мы определяем переменные, которые будут прогнозироваться. Это может быть, например, уровень производства конкретного вида продукции в следующем квартале, прогноз суммы продажи этой продукции и т.д.

При выборе переменных следует учитывать доступность ретроспективных данных, предпочтения лиц, принимающих решения, окончательную стоимость *Data Mining*.

Часто при решении задач *прогнозирования* возникает необходимость предсказания не самой переменной, а изменений ее значений.

Второй вопрос при решении задачи *прогнозирования* - определение следующих параметров:

- периода *прогнозирования* ;
- горизонта *прогнозирования* ;
- интервала *прогнозирования*.

***Период прогнозирования*** - основная единица времени, на которую делается прогноз.

Например, мы хотим узнать доход компании через месяц. Период *прогнозирования* для этой задачи - месяц.

***Горизонт прогнозирования*** - это число периодов в будущем, которые покрывает прогноз.

Если мы хотим узнать прогноз на 12 месяцев вперед с данными по каждому месяцу, то период *прогнозирования* в этой задаче - месяц, горизонт *прогнозирования* - 12 месяцев.

***Интервал прогнозирования*** - частота, с которой делается новый прогноз.

*Интервал прогнозирования* может совпадать с периодом *прогнозирования*.

Рекомендации по выбору параметров *прогнозирования*.

При выборе параметров необходимо учитывать, что горизонт *прогнозирования* должен быть не меньше, чем время, которое необходимо для реализации решения, принятого на основе этого прогноза. Только в этом случае *прогнозирование* будет иметь смысл.

С увеличением горизонта *прогнозирования* точность прогноза, как правило, снижается, а с уменьшением горизонта - повышается.

Мы можем улучшить качество *прогнозирования*, уменьшая время, необходимое на реализацию решения, для которого реализуется прогноз, и, следовательно, уменьшив при этом горизонт и ошибку *прогнозирования*.

При выборе интервала *прогнозирования* следует выбирать между двумя рисками: вовремя не определить изменения в анализируемом процессе и высокой стоимостью прогноза. При длительном интервале *прогнозирования* возникает риск не идентифицировать изменения, произошедшие в процессе, при коротком - возрастают издержки на *прогнозирование*.

При выборе интервала необходимо также учитывать стабильность анализируемого процесса и стоимость проведения прогноза.

**Точность прогноза**

***Точность прогноза***, требуемая для решения конкретной задачи, оказывает большое влияние на прогнозирующую систему. Ошибка прогноза зависит от используемой системы прогноза.

Чем больше ресурсов имеет такая система, тем больше шансов получить более точный прогноз. Однако *прогнозирование* не может полностью уничтожить риски при принятии решений. Поэтому всегда учитывается возможная ошибка *прогнозирования*.

*Точность прогноза* характеризуется ошибкой прогноза.

Наиболее распространенные виды ошибок:

- **Средняя ошибка (СО)**. Она вычисляется простым усреднением ошибок на каждом шаге. Недостаток этого вида ошибки - положительные и отрицательные ошибки аннулируют друг друга.
- **Средняя абсолютная ошибка (САО)**. Она рассчитывается как среднее абсолютных ошибок. Если она равна нулю, то мы имеем совершенный прогноз. В сравнении со средней квадратической ошибкой, эта мера "не придает слишком большого значения" выбросам.
- **Сумма квадратов ошибок (SSE)**, среднеквадратическая ошибка. Она вычисляется как сумма (или среднее) квадратов ошибок. Это наиболее часто используемая оценка точности прогноза.

• **Относительная ошибка (ОО)**. Предыдущие меры использовали действительные значения ошибок. Относительная ошибка выражает качество подгонки в терминах относительных ошибок.

**Виды прогнозов**

Прогноз может быть краткосрочным, среднесрочным и долгосрочным.

**Краткосрочный прогноз** представляет собой прогноз на несколько шагов вперед, т.е. осуществляется построение прогноза не более чем на 3% от объема наблюдений или на 1-3 шага вперед.

**Среднесрочный прогноз** - это прогноз на 3-5% от объема наблюдений, но не более 7-12 шагов вперед; также под этим типом прогноза понимают прогноз на один или половину сезонного цикла. Для построения краткосрочных и среднесрочных прогнозов вполне подходят статистические методы.

**Долгосрочный прогноз** - это прогноз более чем на 5% от объема наблюдений.

При построении данного типа прогнозов статистические методы практически не используются, кроме случаев очень "хороших" рядов, для которых прогноз можно просто "нарисовать".

До сих пор мы рассматривали аспекты *прогнозирования*, так или иначе связанные с процессом принятия решения. Существуют и другие факторы, которые необходимо учитывать при прогнозировании.

Задача 1. Известно, что анализируемый процесс относительно стабилен во времени, изменения происходят медленно, процесс не зависит от внешних факторов.

Задача 2. Анализируемый процесс нестабилен и очень сильно зависит от внешних факторов.

Решение первой задачи должно быть сосредоточено на использовании большого количества ретроспективных данных. При решении второй задачи особое внимание следует обратить на оценки специалиста в предметной области, эксперта, чтобы иметь возможность отразить в *прогнозирующей модели* все необходимые внешние факторы, а также уделить время для сбора данных по этим факторам (сбор внешних данных часто намного сложнее сбора внутренних данных информационной системы). Доступность данных, на основе которых будет осуществляться *прогнозирование*, - важный фактор построения прогнозной модели. Для возможности выполнения качественного прогноза данные должны быть представительными, точными и достоверными.

**Методы прогнозирования**

Методы *Data Mining*, при помощи которых решаются задачи *прогнозирования*, будут рассмотрены во втором разделе курса. Среди распространенных методов *Data Mining*, используемых для *прогнозирования*, отметим нейронные сети и линейную регрессию.

Выбор метода *прогнозирования* зависит от многих факторов, в том числе от параметров *прогнозирования*. Выбор метода следует производить с учетом всех специфических особенностей набора ретроспективных данных и целей, с которыми он строится.

Программное обеспечение *Data Mining*, используемое для *прогнозирования*, должно обеспечивать пользователя точным и достоверным прогнозом. Однако получение такого прогноза зависит не только от программного обеспечения и методов, заложенных в его основу, но также и от других факторов, среди которых полнота и достоверность исходных данных, своевременность и оперативность их пополнения, квалификация пользователя.

**Задача визуализации**

***Визуализация*** - это *инструментарий*, который позволяет увидеть конечный результат вычислений, организовать управление вычислительным процессом и даже вернуться назад к исходным данным, чтобы определить наиболее рациональное направление дальнейшего движения.

С задачей *визуализации* можно подробно ознакомиться по материалам конференций, среди которых, например, *CHI* и ACM-*SIGGraph*, а также в периодической литературе, в частности, по материалам журнала "*IEEE* Trans. *visualization and computer graphics*".

В результате использования *визуализации* создается *графический образ данных*. Применение *визуализации* помогает в процессе анализа данных увидеть аномалии, структуры, *тренды*. При рассмотрении задачи *прогнозирования* мы использовали графическое *представление временного ряда* и увидели, что в нем присутствует *сезонная компонента*. В предыдущей лекции мы рассматривали задачи классификации и кластеризации, и для иллюстрации распределения объектов в двухмерном пространстве также использовали визуализацию.

Можно говорить о том, что применение *визуализации* является более экономичным: линия *тренда* или скопления точек на диаграмме рассеивания позволяет аналитику намного быстрее определить закономерности и прийти к нужному решению. Таким образом, здесь идет речь об использовании в *Data Mining* не символов, а образов.

Главное преимущество *визуализации* - практически полное отсутствие необходимости в специальной подготовке пользователя. При помощи *визуализации* ознакомиться с информацией очень легко, достаточно всего лишь бросить на нее взгляд.

Хотя простейшие виды *визуализации* появились достаточно давно, ее использование сейчас только набирает силу. *Визуализация* не направлена исключительно на совершенствование техники анализа - по словам Скотта Лейбса, в некоторых случаях *визуализация* может даже заменить её.

*Визуализации* данных может быть представлена в виде: графиков, схем, гистограмм, диаграмм и т.д.

Кратко роль *визуализации* можно описать такими ее возможностями:

- поддержка интерактивного и согласованного исследования;
- помощь в представлении результатов;
- использование глаз (зрения), чтобы создавать зрительные образы и осмысливать их.

**Плохая визуализация**

Результаты *визуализации* иногда могут вводить пользователя в заблуждение. Приведем простой пример плохой *визуализации*. Допустим, мы

имеем базу "Прибыль компании А" за период с 2000 по 2005 года, она представлена в табличном виде в таблице 6.1.

Таблица 6.1. Прибыль компании А

| год | прибыль |
|---|---|
| 2000 | 1100 |
| 2001 | 1101 |
| 2002 | 1104 |
| 2003 | 1105 |
| 2004 | 1106 |
| 2005 | 1107 |

Построим гистограмму в Excel по этим данным.

Гистограмма представляет собой визуальное изображение распределения данных.

Эта информация отображается при помощи серии прямоугольников или полос одинаковой ширины, высота которых указывает количество данных в каждом классе.

Используя все значения построения графика, принятые по умолчанию, получаем гистограмму, приведенную на рис. 6.4.

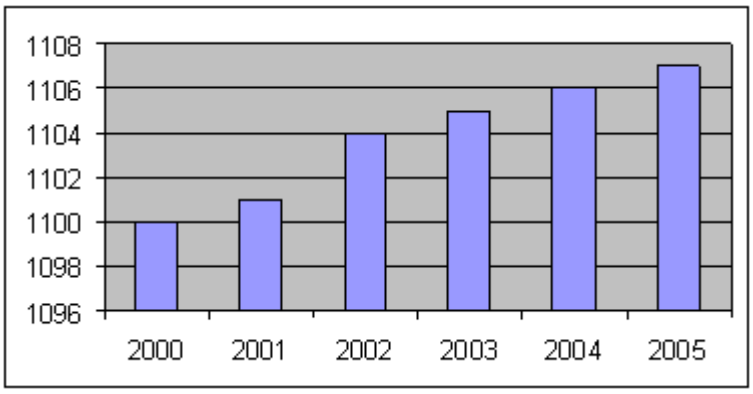


**Рис. 6.4.** Гистограмма, минимальное значение оси у равно 1096

Данный рисунок демонстрирует значительный рост прибыли компании А за период с 2000 по 2005 года. Однако, если мы обратим внимание на ось у, показывающую величину прибыли, то увидим, что эта ось пересекает ось х в значении, равном 1096. Фактически, ось у со значениями от 1096 до 1108 вводит пользователя в заблуждение. Изменив значения параметров, отвечающих за формат оси у, получаем график, приведенный на рис. 6.5.

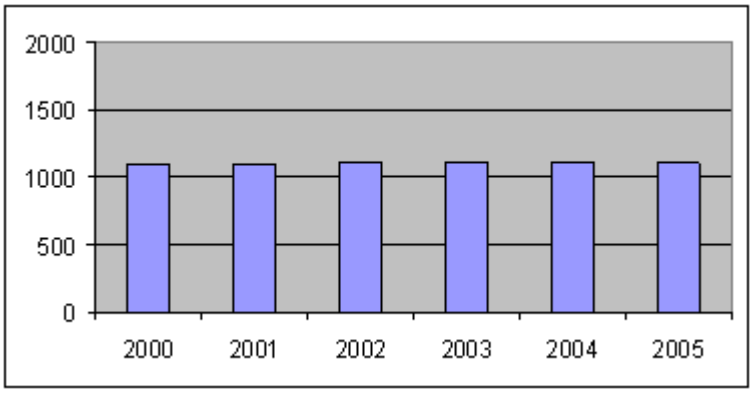


**Рис. 6.5.** Гистограмма, минимальное значение оси y равно 0

Ось y со значениями от 0 до 2000 дает пользователю правильную информацию о незначительном изменении прибыли компании.

Если речь идет о большой размерности и сложности исходных данных, средства *визуализации* обеспечивают их резкое уменьшение, конденсируя, быть может, миллионы записей данных в простые, легкие для понимания и манипулирования представления . Такие представления называют визуальным или графическим способом представления информации. Визуализацию можно считать ключевым фактором в исследовании данных, полученных при помощи инструментов *Data Mining*. В таких случаях говорят о визуальном *Data Mining*.

Методы *визуализации*, среди которых представления информации в одно-, двух-, трехмерном и более измерениях, а также другие способы отображения информации, например, параллельные координаты, "лица Чернова", будут рассмотрены в следующем разделе курса.