МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ
федеральное государственное бюджетное образовательное учреждение
высшего профессионального образования
«УЛЬЯНОВСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ»

ИА

Учебное пособие

Составители:
В. В. Воронина
А. А. Романов

Ульяновск
УлГТУ
2015

ББК 32.973-04я73
УДК 004.3(075.8)
А 76

Рецензенты:
доктор технических наук, профессор И. В. Семушин
кафедра «Информационные технологии» Ульяновского государственного университета

Утверждено редакционно-издательским советом университета в качестве учебного пособия

ИА : учебное пособие / составители В. В. Воронина, А. А. Романов. – Ульяновск: УлГТУ, 2015. – 94 с.

ISBN ??????????

В рамках данного пособия рассматривается одно из направлений обработки данных - Data Mining, а именно интеллектуальный анализ временных рядов. В последние годы к этой тематике проявляется большой интерес поскольку с использованием временных рядов возможно моделировать большое количество самых разнообразных явлений и процессов, которые являются источником рядов. Построенные модели временных рядов призваны выявлять структурные особенности исследуемых процессов с целью их анализа и прогнозирования состояния. Прогноз временного ряда используется для повышения эффективности принятия решений.

Пособие предназначено для студентов направления 09.03.04 «Программная инженерия», изучающих дисциплину «Интеллектуальный анализ данных».

УДК 004.3(075.8)
ББК 32.973-04я73

ISBN ??????????

ОГЛАВЛЕНИЕ

ПРЕДИСЛОВИЕ

В данном пособии представлено описание подхода к разработке программной системы, предназначенной для моделирования временных рядов. Освещены особенности построения научно-исследовательского программного обеспечения.

В главе 1 представлено...

Пособие предназначено для студентов направления 09.03.04 «Программная инженерия», изучающих дисциплину «Интеллектуальный анализ данных»

1 ОСНОВНЫЕ ОПРЕДЕЛЕНИЯ ВРЕМЕННЫХ РЯДОВ

1.1 Определение и свойства временного ряда

Временной ряд определяется как последовательность значений, упорядоченных во времени и характеризующих уровень состояния и изменения наблюдаемого показателя. Особенностью временного ряда является то, что каждое значение показателя зависит от прошлых состояний, т.е. важен порядок следования значений .

Временные ряды различаются по следующим признакам:

1. *Длина ряда*. Здесь употребляется в смысле числа наблюдений параметра, хотя может означать и время, прошедшее от начального до конечного наблюдения. Кендел в [**?**] говорит, что в отличии от обычной статистической работы, в анализе временных рядов количество информации не является пропорциональным числу членов выборки, т.к. «...последовательные величины не являются независимыми».
2. *Дискретность и непрерывность*. Определяется характером изменения времени, в течение которого производится наблюдение. Согласно дискретные ряды могут быть получены из непрерывных двумя способами:
   - выборкой из непрерывных временных рядов через определенный интервал;
   - накоплением значения в течение некоторого периода времени.
3. *Детерминированность*. Если будущие значения временного ряда определены какой-либо математической функцией, тогда временной ряд называется детерминированным. Если же будущие значения описываются только с помощью распределения вероятностей, то временной ряд называется недетерминированным, случайным или стохастическим. В свою очередь стохастический временной ряд может быть стационарным или нестационарным. Ряд называется стационарным, если его свойства не зависят от начала отсчета времени. «В частности, он имеет постоянное математическое ожидание (т.е. среднее значение, относительно которого он варьирует), постоянную дисперсию, определяющую размах его колебаний относительно среднего значения, а также постоянную автоковариацию и коэффициенты автокорреляции». Чтобы дискретный ряд был строго стационарным,

дисперсия любой совокупности наблюдений не должно изменяться при сдвиге времени наблюдения на любое целое число.

4. *Моментные и интервальные* временные ряды. Интервальный ряд - последовательность, в которой уровень явления относят к результату, накопленному за определенный интервал времени. Если же уровень ряда характеризует изучаемое явление в конкретный момент времени, то совокупность уровней образует моментный ряд. Важное отличие моментных рядов от интервальных состоит в том, что сумма уровней интервального ряда дает вполне реальный показатель – общее значение за интервал.
5. *Полные и неполные* временные ряды. Полные ряды имеют место, когда даты регистрации или окончания периодов следуют друг за другом с равными интервалами, неполные - когда принцип равных интервалов не соблюдается.

1.2 Компоненты временного ряда

В изучении временных рядов большое место занимает вопрос о закономерностях их движения на протяжении длительного периода. Выявление зависимостей, действующих во времени является сложной процедурой, поскольку данные зависимости формируются под действием многих факторов. Выделяется две группы факторов:

- определяющие основную тенденцию динамики;
- вызывающие случайные колебания.

Тогда по временной ряд можно представить в следующем виде:

$$x_t = \xi_t + \varepsilon_t$$

где, ε_t генерируется случайным неавтокоррелированным процессом с нулевым математическим ожиданием и конечной (не обязательно постоянной) дисперсией, а величина ξ_t может быть генерирована либо детерминированной функцией, либо случайным процессом, либо какой-нибудь их комбинацией. Величины ε_t и ξ_t различаются характером воздействия на значения последующих членов ряда. Переменная ε_t влияет только на значение синхронного ей члена ряда, в то время как величина ξ_t в известной

степени определяет значение нескольких или всех последующих членов ряда. Основной тенденцией, или трендом, называется характеристика процесса изменения явления за длительное время, освобожденная от случайных колебаний, создаваемых второй группой факторов. В модели тренд обозначается через ξ_t и может быть выражен как детерминированной так и случайной функциями или их комбинацией. Колеблемостью следует называть отклонения уровней отдельных периодов времени от тенденции динамики (тренда).

Однако можно наблюдать иерархию тенденций и колебаний: та величина, которая для, например столетия, выступает как колебания, на интервале времени низшего порядка, например трех-пяти лет, может выступать как тенденция.

Временной ряд следует рассматривать как смесь четырех компонент:

- тренда,
- регулярных колебаний относительно тренда,
- сезонной компоненты,
- остатка или несистематического случайного эффекта.

В общем случае временной ряд представляется в виде суммы этих 4-х компонент. Одной из задач анализа временных рядов является разложение ряда на составляющие его компоненты с целью их изучения. Компоненты временного ряда ненаблюдаемы. Они являются теоретическими величинами. Их выделение и составляет предмет анализа временного ряда.

1.3 Классификация задач моделирования и анализа временных рядов

Основным средством анализа и прогноза временного ряда является модель. Модели временных рядов удобно применять, в частности, для дискретных систем «...т.е. таких систем, в которых возможность произвести наблюдение и предпринять регулирующие действия возникает через равные интервалы времени времени». Понятие модель используется в двух значениях: как модель временного ряда, выражающая закон генерирования членов ряда, и как прогнозная модель. Главное отличие

этих двух типов моделей в том, что на выходе модели временного ряда фактические члены ряда, а на выходе прогнозной модели — оценки будущих членов ряда. Также выделяются следующие важные прикладные области применения моделей временных рядов:

1. Прогнозирование будущих значений временного ряда по его текущим и прошлым значениям.
2. Определение передаточной функции системы.
3. Проектирование простых регулирующих схем с прямой и обратной связями.

Выделяются следующие цели анализа временных рядов:

1. Построение системы математического вида, которая описывает поведение временного ряда в сжатом виде.
2. Для объяснения поведения временного ряда с помощью других переменных в качестве гипотезы строится модель.
3. Результаты анализа, полученные в 1 или 2 могут быть использованы для прогнозирования поведения ряда.
4. В случае 2 возможен контроль системы путем выработки сигналов о наступающих изменениях или путем исследования того, что может случиться, если изменить некоторые из параметров модели.
5. Анализ совместного развития во времени нескольких переменных.

Объединяя эти точки зрения можно выделить следующие задачи, которые могут быть решены при моделировании временных рядов:

1. Построение формализованного представления моделируемой системы с выделением ее значимых параметров - определение природы временного ряда.
2. Прогнозирование будущих значений временного ряда, т.е. определение вида передаточной функции.

Прогнозирование – это научное, основанное на системе установленных причинно-следственных связей и закономерностей, выявление состояния и вероятностных путей развития явлений и процессов.

Выделяются следующие типы прогнозов:

1. В зависимости от целей исследования прогнозы делятся на поисковые и нормативные:

- Нормативный прогноз – это прогноз, который предназначен для указания возможных путей и сроков достижения заданного, желаемого конечного состояния прогнозируемого объекта, то есть нормативный прогноз разрабатывается на базе заранее определенных целей и задач.
- Поисковый прогноз не ориентируется на заданную цель, а рассматривает возможные направления будущего развития прогнозируемого объекта, то есть выявление того, как будет развиваться объект в будущем полностью зависит от сохранения существующих тенденций.

2. В зависимости от специфики области применения прогноза и от объекта прогнозирования прогнозы подразделяются на:
   - естественноведческие – это прогнозы в области биологии, медицины и так далее;
   - научно-технические - это, например, инженерное прогнозирование технических характеристик узлов, деталей и так далее.
3. В зависимости от масштабности объекта, прогнозы бывают:
   - глобальные – рассматривают наиболее общие тенденции и закономерности в мировом масштабе;
   - макроэкономические – анализируют наиболее общие тенденции явлений и процессов в масштабе экономики страны в целом;
   - структурные (межотраслевые и межрегиональные) – предсказывают развитие экономики в разрезе отраслей;
   - региональные – предсказывают развитие отдельных регионов;
   - отраслевые – прогнозируют развитие отраслей;
   - микроэкономические – предсказывают развитие отдельных предприятий, производств и так далее.
4. По сложности прогнозы различают:
   - сверхпростые – прогноз на основе одномерных временных рядов, когда отсутствуют связи между признаками;
   - простые – прогнозы, предполагающие учет оценки связей между факторными признаками;

- сложные – прогнозы, оценка связей между признаками в которых определяется на основе системы уравнений или многофакторного динамического прогнозирования.

Модели основаны на допущении о том, что основные факторы и тенденции прошлого периода сохранятся и на период прогноза, или что направление и изменение тенденций в рассматриваемой перспективе можно обосновать и учесть, т.е. предполагается большая инерционность систем.

2 МЕТОДЫ МОДЕЛИРОВАНИЯ ВРЕМЕННЫХ РЯДОВ

2.1 Статистические методы и средства анализа рядов

Для прогнозирования будущих показателей имеющегося временного ряда необходимо построить модель ряда, которая наиболее полно отражает изменение исследуемого ряда. Существует множество моделей, описывающих различные стохастические процессы и имеющие различные формы представления, но среди них выделяют следующие, имеющие наибольшую ценность в практике:

- авторегрессионные модели (первого порядка, второго порядка)- (Autoregressive Moving Average, ARMA);

$$y(t) = c + \sum_{i=1}^{p} a_i y(t-i) + \varepsilon_t$$

- модели скользящего среднего (Moving Average);

$$\bar{y}(k) = \frac{1}{n} \sum_{t=k}^{n+k} y(t)$$

- интегральные модели (Autoregressive Integrated Moving Average, ARIMA);

$$\Delta^d y(t) = c + \sum_{i=1}^{p} a_i \Delta^d y(t-i) + \frac{1}{q} \sum_{j=1}^{q} b_j \Delta^d y(t-j) + \varepsilon_t$$

В целом, статистический подход к анализу временных рядов заключается в выявлении и моделировании его детерминированных компонент на основе аддитивной (или мультипликативной) параметрической функциональной модели, приведении остатков к стационарному виду, при моделировании которых полученные ошибки удовлетворяли ограничениям модели

Изучение временных рядов осуществляется при помощи вероятностно-статистических моделей. При этом выделяют характеристики временного ряда $x(t)$:

- математическое ожидание $a(t) = Mx(t)$;
- дисперсия $\sigma^2(t) = Dx(t)$;
- автокорреляционая функция временного ряда: $p(t,s) = \frac{M(x(t)-a(t))(x(s)-a(s))}{\sigma(t)\sigma(s)}$.

В зависимости от значений этих показателей временные ряды делятся на стационарные и нестационарные.

Стационарные временные ряды можно выделить с помощью такого критерия как неизменность ранее перечисленных характеристик временного ряда: математическое ожидание и дисперсия являются постоянными величинами, автокорреляционная функция зависит только от разности $t - s$. Иначе временной ряд является нестационарным.

Нестационарные временные ряды. В терминах статистики в поведении временного ряда обычно выявляют две основные тенденции – тренд и периодические колебания. При анализе временных рядов стараются выделить тренд. Если затем вычесть его из исходных данных то остается колеблющийся ряд - случайные скачки, нерегулярности. Существенным понятием для тренда является гладкость. Представляются следующие методы представления тренда:

- функциональная форма, полином низкой степени;
  плюсы: простая форма представления;
  недостатки:
  - трудно производить обновление,
  - сложно оценивать параметры функции,
  - подбор новой функции после добавления новых точек ряда может исказить полученную ранее последовательность.
- скользящие средние;
  плюсы: упрощенная форма полиномиального представления;
  минусы: возможность сгладить циклическую, краткосрочную составляющую тренда, запаздывание значений относительно исходного ряда;
- взвешенные скользящие средние;
  плюсы: учитывается расстояние от точки до середины интервала сглаживания;
  минусы: усиливает зависимость уровней ряда друг от друга;
- экспоненциальная средняя;
  плюсы: влияние прошлых наблюдений затухает по мере удаления от момента, для которого определяется средняя;
  минусы: влияние будет значительным для первых членов ряда, следовательно ряд должен иметь достаточно большое количество уровней;

условия применения: наличие предыдущего значения в связи с рекурентным процессом вычисления; выбор оптимальной постоянной сглаживания, характеризующей скорость реакции модели на изменение уровней.

Главный недостаток этих методов в том, что они рассматривают временной ряд изолированно от других явлений, и если даже имеется дополнительная информация, она может быть использована исследователем лишь путем регулирования скорости адаптации. Кроме того, точность прогнозов заметно падает при долгосрочном прогнозировании.

Следующим шагом после сглаживания тренда является моделирование стационарного процесса - разности между трендом и временным рядом. Такие ряды составляют другой класс временных рядов (в отличие от рядов, для которых возможно применить сглаживание с помощью скользящего среднего с конечным отрезком усреднения) , известный как класс авторегрессий. «К данному ряду можно относиться как к генерируемому механизмом, в котором значение ряда в момент времени t выражается через прошлые значения - систематическая зависимость от прошлой истории плюс случайная погрешность».

Минусы стационарных моделей:

Не существуют однозначных и эффективных критериев и методов определения факта наличия детерминированного тренда. Существуют статистические критерии проверки гипотезы о наличии тренда. Но эти критерии используют двухальтернативный базис: тренд или случайная компонента (метод восходящих-нисходящих серий), тренд или периодическая компонента, регулярная или случайная компонента.

Статистические модели характеризуются невысоким качеством при моделировании коротких временных рядов (количество наблюдений меньше 40)

Минусы статистических методов:

- отсутствие в модели представлений о структуре и системе связей реального объекта, что вносит субъективизм в выбор как самой модели, так и ее структуры;
- трудность построения моделей при условии, что данные хранятся в разных временных рядах и (или) имеют временные сдвиги относительно друг друга;

- значительная чувствительность получаемых результатов к недостатку информации и (или) ее зашумленности;
- потребность в высокой квалификации математиков-программистов;
- зависимость результата прогноза от квалификации аналитика в конкретной предметной области.

2.2 Нейросетевые методы и средства анализа рядов

Последняя тенденция - сочетание линейных и нелинейных моделей для прогнозирования временных рядов является активной областью исследований.

Гибкость и силовые возможности нейронных сетей применительно к распознаванию сделали их привлекательной альтернативой, когда структура данных порождающей системы неизвестна. Однако, если начать формулировать прогностическую модель, ИНС, как правило, трудно интерпретировать и для проверки статистической значимости параметров.
.

Нейронные сети благодаря своим свойствам хорошо зарекомендовали себя при обработке данных. Одной из важных особенностей является способность к обучению и обобщению накопленных знаний. На ограниченном множестве данных сеть, обобщая полученную информацию, показывает хорошие результаты на данных, которые не использовались при обучении. Функции, выполняемые сетями:

- аппроксимация,
- классификация и распознавание образов,
- прогнозирование,
- идентификация и оценивание,
- ассоциативное управление.

«В каждом из названных приложений нейронная сеть играет роль универсального аппроксиматора функции от нескольких переменных, реализуя нелинейную функцию $y = f(x)$, где x - входной вектор, y - реализация векторной функции нескольких переменных.» Постановки значительного количества задач моделирования могут быть сведены именно к такому представлению. «Для классификации и распознавания образов сеть обучается важнейшим их признакам, таким, как геометрическое отображение конечной структуры изображения, относительное

расположение важнейших элементов образа, компоненты преобразования Фурье и другие подобные факторы. В процессе обучения выделяются признаки, отличающие образы друг от друга, которые и составляют базу для принятия решений об отнесении образов к соответствующим классам.»

«При решении задач прогнозирования роль нейронной сети состоит в предсказании будущей реакции системы по ее предшествующему поведению. Обладая информацией о значениях переменной x в моменты, предшествующие прогнозированию $x(k-1), x(k-2), ..., x(k-N)$, сеть вырабатывает решение, каким будет наиболее вероятное значение последовательности $\widehat{x}(k)$ в текущий момент k».

«Моделирование ВР в рамках нейросетевого подхода сводится к задаче наилучшей аппроксимации нелинейной функции от многих переменных по набору примеров, заданных историей временного ряда:

$$\widehat{y}_{k+1} = \phi(y_k, ..., y_{k-n+1}) + \varepsilon_{k+1}$$

где $\widehat{y}_{k+1}$ - прогнозируемое значение уровня временного ряда;
$y_k, ..., y_{k-n+1}$ - наблюдаемые значения уровней временного ряда;
$\phi(y_k, ..., y_{k-n+1})$ - некоторая нелинейная функция, параметрической моделью которой служит нейронная сеть; ε_{k+1} - ошибка прогноза; n - порядок модели».

Эффективность нейронной сети во многом зависит от ее структуры. Взаимодействие между различными узлами сети задаются с помощью структуры. Структура ИНС не является уникальной для данной задачи, и могут существовать различные способы определить структуру, соответствующую проблеме. В зависимости от задачи, может оказаться целесообразным иметь больше одного скрытого слоя, упреждающие или обратные связи, а в некоторых случаях, прямые связи между входным и выходным слоем.

Успех нейросетевого моделирования зависит от

- типа данных;
- умения аналитика в выборе подходящей нейросетевой модели и / или;
- численных методов, используемых в модели и для вычисления прогнозов.

Исследования показали, что хорошая модель нейросетевая модель для временных рядов должна быть выбрана путем сочетания традиционного

моделирования со знанием анализа временных рядов и проблем, связанных с параметрами нейросетевых моделей.

Влияние на качество нейросетевой модели оказывает также размер окна - количество данных, участвующих в прогнозе. Если окно слишком мало, то аттрактор системы проецируется на пространство недостаточной размерности. Кроме того, окно слишком большого размера может привести к проблемам: кроме основной информации в окно попадает шум.

Выделяется три сущности, необходимые для построения нейронной сети:

1. Модель нейронной сети, архитектуры.
2. Алгоритм обучения, который определяет веса связей.
3. Функция, которая определяет выход каждого нейрона, функция активации.

Модели нейронной сети могут быть разделены на следующие три типа:

1. Сети с прямой связью: рассматривают восприятие моделью обратного распространения, в основном используется в таких областях, как прогнозирование и распознавания образов;
2. Сети с обратной связью: в основном используется для ассоциативной памяти и оптимизации расчетов;
3. Самоорганизующиеся сети: рассматривают модели адаптивной резонансной теории и модели Кохонена, используется для кластерного анализа

Ниже в таблице 2.1 мы перечислим сопоставимые свойства, взятые из теории нейронных сетей и статистических методов:

В настоящее время нейронные сети наиболее часто используются при интеллектуальном анализе данных временных рядов. Преимущества применения нейронных сетей:

- высокая точность: нейронные сети могут аппроксимировать сложные нелинейные отображения;
- устойчивость к шуму: нейронные сети являются очень гибкими по отношению к неполным, пропавшим и зашумленным данным;
- независимость от предыдущих предположений: нейронные сети не делают априорных предположений о распределении данных или формы взаимодействия между факторами;

Таблица 2.1

Нейронная сеть	**Статистика**
Функции:	переменные
входы	независимые переменные
выходы	спрогнозированное значение
цели или значения для обучения	зависимые переменные
ошибки	остатки
обучение	расчет
функция ошибок	оценочный критерий
образцы или пары обучения	наблюдения
(синаптические) веса	оценки параметров
нейроны высокого порядка	взаимодействия
функциональные связи	преобразования
обучения с учителем	регрессии и дискриминантный анализ
неконтролируемое обучение	сжатие данных
конкурентного обучения или кластерный анализ	адаптивный вектор квантования
обобщение	интерполяции или экстраполяции

- простота обслуживания: нейронные сети могут быть использованы с новыми данными, что делает их полезными в динамических средах;
- могут быть реализованы в параллельном оборудовании;
- когда элемент нейронной сети выходит из строя, она может продолжать работать без каких-либо проблем.

Возникающие проблемы:

- нет общих методов для определения оптимального количества нейронов, необходимого для решения любой задачи;
- трудно выбрать набор обучающих данных, который будет достаточен для решения задачи.

Искусственные нейронные сети имеют как общие проблемы, такие как сходимость, устойчивость, наличие минимальных параметров настройки, так и частные проблемы анализа временных рядов: малая скорость обучения, попадание в локальный минимум, трудность определения параметров тренировки. Объединение генетических алгоритмов с нейронными сетями помогают достичь более высоких результатов.

2.3 Методы и средства интеллектуального анализа данных

Интеллектуальный анализ данных процесса может быть составлен на три основных этапа: подготовка данных, интеллектуального анализа данных, выражения и интерпретации результатов. Интеллектуальный анализ данных это итерационный процесс повторения этих трех фаз. Детали показаны на рис. 2.1.

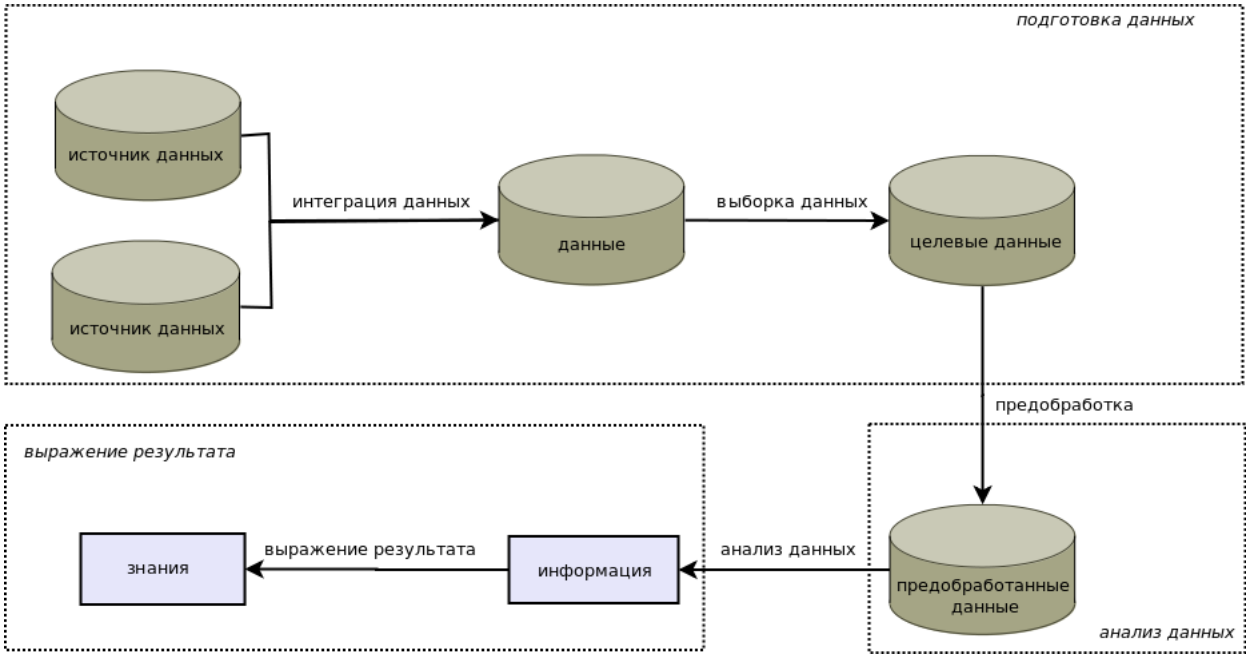

Рис. 2.1. Основной процесс анализа данных

Интеллектуальный анализ данных на основе нейронных сетей состоит из подготовки данных, правила извлечения и оценивание правил, как показано на рис. 2.2

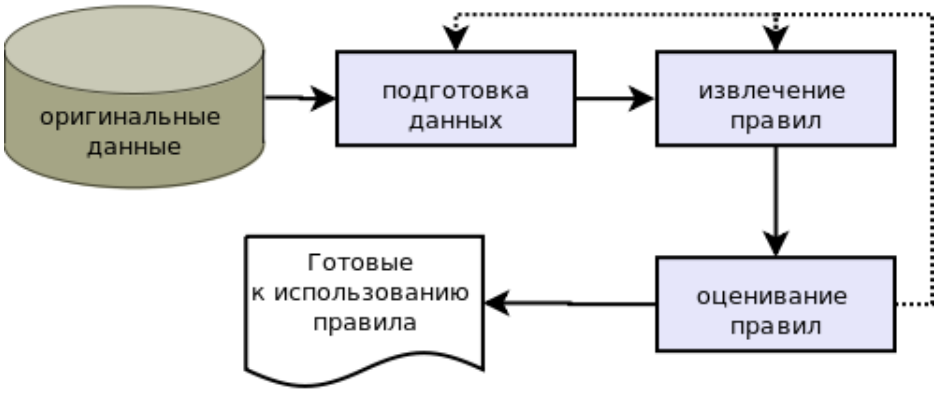

Рис. 2.2. Процесс анализа данных, основанный на нейронной сети

1. Подготовка данных. Подготовка данных является определение и процесс интеллектуального анализа данных, чтобы сделать его пригодным конкретного интеллектуального анализа данных методом. Подготовка данных является первым важным шагом в интеллектуальном анализе данных и играет решающую роль во всем процессе интеллектуального анализа данных. В основном включает в себя следующие процессы:
   (a) Очистка данных. Устранение шумов данных и устранение несоответствия в данных.
   (b) Выборка данных.
   (c) Предварительная обработка.
   (d) Выражение данных. Выражение данных состоит в преобразовании после предварительной обработки в форму, которая может быть принята алгоритмом интеллектуального анализа данных, основанном на нейронной сети. Интеллектуальный анализ данных на основе нейронной сети может работать только с числовыми данными, так что это необходимо трансформировать символьные данные в цифровые. Простейший метод заключается в создании таблицы с взаимно-однозначным соответствием между символьными данными и числовыми.
2. Извлечение правил;
3. Оценивание правил.

Хотя цель оценки правил зависит от каждого конкретного применения, но, в общем случае, правила могут оцениваться по следующим критериям.

- найти оптимальную последовательность извлечения правил, что позволяет получать лучшие результаты на данном наборе данных;
- проверить правильность извлечения правил;
- определить, сколько знаний в нейронной сети не были извлечены;
- обнаружить несоответствия между извлеченными правилами и обученной нейронной сетью.

Есть еще более серьезный принципиальный недостаток статистических пакетов, ограничивающий их применение в Data Mining. Большинство методов, входящих в состав пакетов опираются на статистическую парадигму, в которой главными фигурантами служат усредненные характеристики выборки. А эти характеристики, как указывалось выше, при

исследовании реальных сложных жизненных феноменов часто являются фиктивными величинами.

Существуют различные варианты применения нейронных сетей для анализа данных такие как интеллектуальный анализ данных на основе самоорганизующихся нейронных сетей и на основе нечеткой нейронной сети. Самоорганизационный процесс - процесс обучения обучения без учителей. Исследуются важные характеристики или некоторые врожденные знания в группе данных, таких как характеристики распределения или кластеризации в соответствии с определенной функцией;

Хотя нейронная сеть имеет сильные функции обучения, классификации, ассоциативной памяти, но в использовании нейронных сетей для интеллектуального анализа данных наибольшую трудность составляет то, что результаты на выходе не могут быть интуитивно понятны. После введение нечеткой функции обработки в нейронную сеть, она позволяет не только увеличить емкость продукционных выражений, но также система становится более стабильной. Главным отличием от обычной нейронной сети является то, что в традиционной нейронной сети образцы могли принадлежать одной категории. Нечеткие сети имеют способность отражать степень принадлежности.

Основные методы и подходы реализации:

- сочетание нейронной сети и интеллектуального анализа данных;
- сочетание обработки знаний и нейронных вычислений; Для оценки реализации алгоритма интеллектуального анализа данных могут быть использованы следующие показатели и характеристики:
  - качество моделирования в условиях шума и сырых данных;
  - модель должна быть понята пользователю и может быть использована для принятия решений;
  - модель может использовать знания для улучшения качества моделирования. Нейронные сети на самом деле можно рассматривать как черный ящик для пользователей, что делает процесс классификации и прогнозирования не понятым для пользователей и невозможным для непосредственного использования для принятия решений.

Методы интеллектуального анализа данных. Классические методами приобретения знаний из массивов данных являются статистические

методы. В интеллектуальном анализе данных используются новые методы, кроме статистических. Эти методы имеют свое происхождение в области искусственного интеллекта. Они отличаются от классических статистических методов в следующим:

- ищут неизвестные и неожиданные отношения, которые могут быть обнаружены путем изучения данных в базе данных. Они пытаются найти в наборах данных такие закономерности, которые принесут пользователям новый взгляд в сфере интересов и позволят ему сформулировать новые гипотезы. Статистические методы используют несколько иной путь. Они проверяют или отвергают гипотезы заявленных априори.
- Методы интеллектуального анализа данных можно использовать и в таких случаях, в которых использование классических статистических методов не подходит. Например, когда имеется большой объем многомерных данных или когда не представляется возможным полагать, что данные имеют некоторое стандартное вероятностное распределение.

В интеллектуальном анализе данных изучаются и используются следующие методы получения знаний:

- статистические методы (прогноз временных рядов, кластерный анализ и др.);
- продукционные правила ЕСЛИ. . . ТО;
- деревья решений;
- генетические алгоритмы;
- нейронные сети.

Продукционные правила образуют базу знаний экспертной системы с продукционной архитектурой системы. При проектировании экспертной системы разработка продукционных правил является результатом обсуждения между инженером и группой экспертов.

В интеллектуальном анализе данных изучаются методы автоматической разработки продукционных правил. Такие методы в основном разработаны для создания правил называются ассоциативными правилами. Целью правил ассоциации является выявление отношений между данными в больших базах данных, позволяют найти элементы, которые предполагают наличие других элементов этой же серии. Проблемы автоматического

создания ассоциативных правил были интенсивно изучались в последнее десятилетие. Были разработан эффективные алгоритмы.

Дерево решений представляет возможным представление решения функцией. Оно используется, когда полное знание данных не является необходимым для принятия соответствующего решения, и когда процесс получения данных является затратным. Существуют алгоритмы для автоматического построения деревьев решений. Автоматическое построение деревьев решений является традиционной частью искусственного интеллекта.

Деревья решения являются одним из наиболее популярных подходов к решению задач Data Mining. Они создают иерархическую структуру классифицирующих правил типа «ЕСЛИ... ТО...» (if-then), имеющую вид дерева. Для принятия решения, к какому классу отнести некоторый объект или ситуацию, требуется ответить на вопросы, стоящие в узлах этого дерева, начиная с его корня. Вопросы имеют вид «значение параметра А больше х?». Если ответ положительный, осуществляется переход к правому узлу следующего уровня, если отрицательный - то к левому узлу; затем снова следует вопрос, связанный с соответствующим узлом.

Популярность подхода связана как бы с наглядностью и понятностью. Но деревья решений принципиально не способны находить «лучшие» (наиболее полные и точные) правила в данных. Они реализуют наивный принцип последовательного просмотра признаков и «цепляют» фактически осколки настоящих закономерностей, создавая лишь иллюзию логического вывода.

Генетический алгоритм является универсальным методом для создания объектов с заданными свойствами. Объекты описаны последовательностью символов, которая по своей форме аналогична геному живых организмов. Генетический алгоритм создает поколения объектов. Новое поколение возникает в результате мутации, кроссовера или их комбинация из этих объектов существующего поколения, которые имеют наибольшее значение фитнесс-функции. Эволюция моделируется генетическим алгоритмом, направленным на генерацию объектов с большим значением функция пригодности или, другими словами, объектов с заданными свойствами.

Генетические алгоритмы удобны тем, что их легко распараллеливать. Например, можно разбить поколение на несколько групп и работать с каждой из них независимо, обмениваясь время от времени несколькими хромосомами. Существуют также и другие методы распараллеливания генетических алгоритмов.

Генетические алгоритмы имеют ряд недостатков. Критерий отбора хромосом и используемые процедуры являются эвристическими и далеко не гарантируют нахождения «лучшего» решения. Как и в реальной жизни, эволюцию может «заклинить» на какой-либо непродуктивной ветви. И, наоборот, можно привести примеры, как два неперспективных родителя, которые будут исключены из эволюции генетическим алгоритмом, оказываются способными произвести высокоэффективного потомка. Это особенно становится заметно при решении высокоразмерных задач со сложными внутренними связями.

Самоорганизующиеся карты Кохонена стали многообещающим метод в кластерном анализе. Они адаптированы для обучения без учителя. Неконтролируемый процесса обучения в SOM может быть кратко описать следующим образом 2.3. После процесса обучения, подобные наборы элементов активируют тот же нейрон. SOM делит входной набор на подмножества подобных записей. Таким образом, SOM представляет собой метод кластерного анализа и часто используется для удобного представления данных.

В интеллектуальном анализе данных самоорганизующиеся карты Кохонена имеют следующие преимущества по сравнению со стандартными статистическими методами:

- Интеллектуальный анализ данных обычно имеет дело с многомерными данными. Запись в базе данных обычно состоит из большого количества элементов. Эти данные не имеют регулярного многомерного распределения и, следовательно, традиционные статистические методы имеют свои ограничения, не являются эффективными. SOM позволяет эффективно обрабатывать многомерные данные.
- Самоорганизующиеся карты предоставляют средства для визуализации многомерных данных.

Основным недостатком нейросетевой парадигмы является необходимость иметь очень большой объем обучающей выборки. Другой существен-

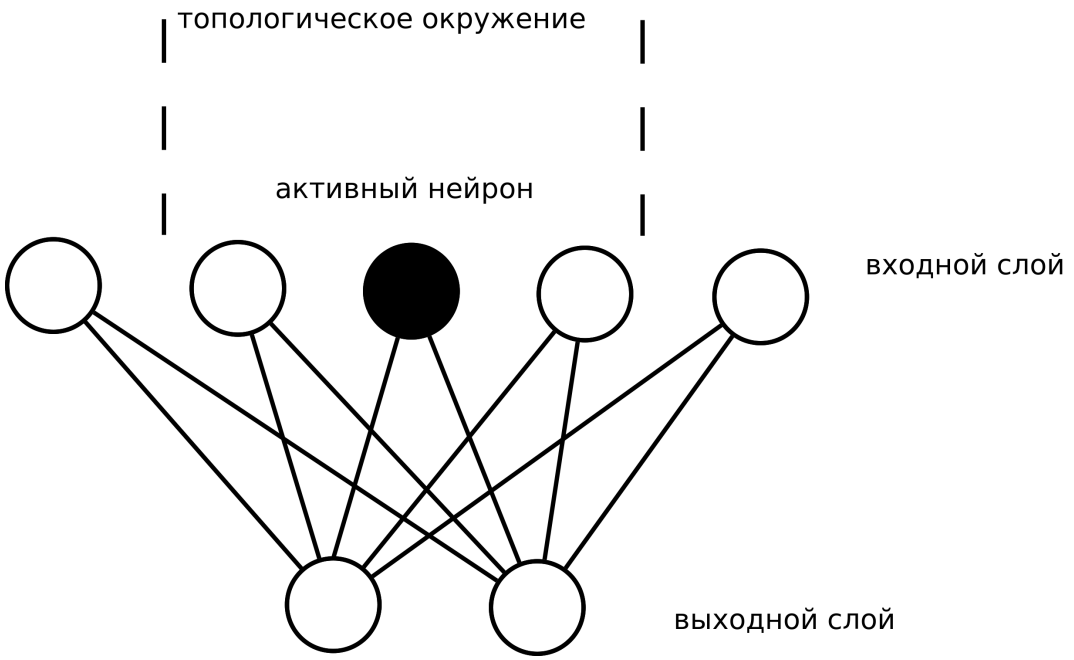

Рис. 2.3. Самоорганизующиеся карты кохонена

ный недостаток заключается в том, что даже натренированная нейронная сеть представляет собой черный ящик. Знания, зафиксированные как веса нескольких сотен межнейронных связей, совершенно не поддаются анализу и интерпретации человеком.

Анализ знаний - нетривиальный процесс. В настоящее время используется множество различных методов интеллектуального анализа данных. Методы анализа данных играют важную роль в управлении сложными системами, имеют большую область применения в экономике. Программное обеспечение, реализующее алгоритмы интеллектуального анализа данных, существуют на рынке, но в настоящее время цена на него высока.

2.4 Методы анализа рядов на основе нечетких систем

Под нечетким временным рядом (НВР) будем понимать упорядоченную последовательность наблюдений над некоторым явлением, состояния которого изменяются во времени, если значение состояния в момент t_i выражено с помощью нечеткой метки $\tilde{x_i} \in \tilde{X}, i \in [1,n]$, n – количество членов ряда, то есть нечеткий временной ряд представим в виде

$$\tilde{Y} = \{t_i, \tilde{x_i}\} \tag{2.1}$$

где $\tilde{x}_i$ – i-тое нечеткое множество(нечеткая метка), t_i – i-тое значение момента времени $t_1 \le t_i \le t_n$, n – количество членов НВР.

Нечеткая метка – это понятие на естественном языке, получаемое посредством фаззификации исходного четкого временного ряда. Таким образом, для каждому четкому элементу временного ряда соответствует пара (X, μ), где X – лингвистическое понятие, μ – степень принадлежности элемента указанному понятию.

Для описания развития моделируемого процесса в лингвистических терминах введем понятие временного ряда нечетких тенденций. Выделим далее базовые операции обработки нечетких тенденций.

«Нечеткой тенденцией (НТ) нечеткого временного ряда будем называть нечеткую метку, выражающую характер изменения (систематическое движение) последовательности нечетких уровней НВР в заданном интервале времени. Нечеткая тенденция выражает поведение НВР в лингвистическом виде, например: ’Рост’, ’Падение’, ’Стабилизация’, ’Колебания’, ’Хаос’. Для нечетких термов, обозначающих тенденцию, возможно применение модификаторов ’очень’, ’более-менее’ и т. д. Отметим важное свойство нечетких экспертных оценок, обусловленное возможностью их ранжирования, что позволяет представить их совокупность в виде некоторой системы (шкалы) с отношениями. Бинарные отношения, образованные на множестве нечетких экспертных оценок, порождают сравнительные оценки по различным критериям, такие, как ’Больше’, ’Меньше’, ’Примерно Равны’, ’Рост’, ’Падение’, ’Предпочтительнее’, ’Лучше’. Такие сравнительные оценки представляют изменения (различие) нечетких меток в различных пространствах: в пространстве объектов, во временном пространстве, в пространстве задач и характеризуют тенденции. Изменения нечетких меток во временном пространстве порождают не- четкий временной ряд с нечеткой тенденцией»

«В соответствии с логикой оценивания будем считать оценку нечеткого уровня ВР - абсолютной нечеткой оценкой, а оценку изменения нечетких уровней (нечеткую тенденцию) - сравнительной нечеткой оценкой».

В качестве инструмента как абсолютного, так и сравнительного нечеткого оценивания Афанасьевой Т.В. была предложена специальная лингвистическая шкала - ACL-шкала (Absolute and Comparative Linguistic)

[Афанасьева, 2008a]. «Абсолютные оценки, полученные по ACL-шкале, соответствуют нечетким оценкам (меткам) уровней нечеткого временного ряда, а сравнительные оценки – нечетким тенденциям НВР».

В отличие от традиционного ВР значениями нечеткого ВР являются нечеткие множества, а не действительные числа наблюдений. В 1993 году Сонг и Чиссом(Song, Chissom) предложили модели стационарных и нестационарных (time-invariant и time-variant) нечетких временных рядов первого порядка (fist-order) и применили разработанные модели для прогнозирования количества регистрирующихся студентов университета штата Алабама, фаззифицировав предварительно четкий временной ряд. Это было первое определение моделей нечетких временных рядов.

Пусть $X_t,(t=1,2,...) \subset R^1$ - универсум, на котором определены нечеткие множества $y_t^i,(i=1,2,...)$ и Y_t - коллекция $y_t^i,(i=1,2,...)$.Тогда $Y_t,(1,2,...)$ называется нечетким ВР.

На практике в большинстве ВР последовательные наблюдения зависимы, так, что:

$$R = \{(y_t, y_{t-1}),(y_{t-1}, y_{t-2}),...\} \subseteq Y_t \times Y_{t-1} \tag{2.2}$$

где Y_t, Y_{t-1} обозначает переменные, а y_t, y_{t-1} - наблюдаемые значения этих переменных. Наиболее частой моделью зависимости является явная функция отображения:

$$f : Y_{t-1} \to Y_t \tag{2.3}$$

представленная линейной функцией (марковским процессом, модель AR):

$$y_t = f(y_{t-1}, \phi, \varepsilon) = \phi y_{t-1} + \varepsilon \tag{2.4}$$

где ε - случайная ошибка, шум. В случае нечеткого ВР в качестве модели авторегрессии используется нечеткое разностное уравнение:

$$y_t^j = y_{t-1}^i \circ R_{ij}(t, t-1) \tag{2.5}$$

$y_t^i \in Y_t, y_{t-1}^i \in Y_{t-1}, i \in I, j \in J, \circ - maxmin$,композиция,

$$R(t, t-1) = \bigcup_{ij} R_{ij}(t, t-1) \tag{2.6}$$

есть система нечетких отношений, которая символически может быть записана в виде $Y_t \to Y_{t-1}$. Систему отношений R в выражении

$$Y_t = Y_{t-1} \circ R(t, t-1) \tag{2.7}$$

называют моделью нечеткого ВР первого порядка, данная модель – важный частный случай общей модели порядка p:

$$Y_t = (Y_{t-1} \times Y_{t-2} \times ... \times Y_{t-p}) \circ R(t, t-p), \tag{2.8}$$

$$R(t, t-p) = \max_{p} \left\{ \min_{j, i_1, i_2, ..., i_p} \{ y_t^j, y_{t-1}^{i_1}, ..., y_{t-p}^{i_p} \} \right\}$$

Моделирование нечетких временных рядов (по Сонгу) состоит в реализации следующих шагов:

1. Фаззификация входных данных – разбиение данных на множество интервалов, определение лингвистических значений нечетких множеств и функций принадлежности.
2. Формирование логических отношений $Y_t \to Y_{t-1}$ и вычисление $R_{ij}(t, t-1)$ для каждой импликации.
3. Вычисление результирующего отношения R как объединение $\bigcup_{i,j} R_{ij}(t, t-1)$
4. Применение полученной модели к входным данным и получение выходных результатов
5. Дефаззификация.

Предложенная Сонгом модель НВР имеет следующие недостатки:

1. Требуется большое количество вычислений для определения нечеткого отношения.
2. Для определения нечеткой максиминной композиции модели, требуется большое количество вычислений, особенно, когда нечеткое отношение очень велико.
3. Точность предсказания не достаточна.
4. Отсутствие четких рекомендаций по выбору количества нечетких множеств, и определению интервалов их носителей. Данные задачи

выполняются экспертом, и, как показывают исследования, от выбора интервалов сильно зависит результат исследования.

5. Проблема длин интервалов: различные длины интервалов могут привести к различным нечетким отношениям, и в свою очередь породить различные модели и результаты прогноза.

В задачах анализа тенденций НВР недостатки данной модели уменьшаются за счет большей размытости исходных данных и, следовательно, пониженным требованиям к точности.

3 МЕТОДИКИ ОЦЕНКИ КАЧЕСТВА ПРОГНОЗОВ ВРЕМЕННЫХ РЯДОВ

Кроссвалидация http://www.long-short.ru/post/kross-validatsiya-cross-validation-304 Проверка на тестовом интевале

4 РАЗРАБОТКА МЕТОДА ПРОГНОЗИРОВАНИЯ ВРЕМЕННЫХ РЯДОВ

рассмотреть общую концепцию: - предобработка - нормализация - моделирование одним или несколькими методами (с декомпозицией на компоненты или без) - оценка качества сформированной модели - формирование прогноза ряда - денормализация

описать несколько вариантов

5 РАЗРАБОТКА АРХИТЕКТУРЫ ПРОГРАММНОЙ СИСТЕМЫ ДЛЯ МЕТОДА ПРОГНОЗИРОВАНИЯ

6 СОЗДАНИЕ СЕРВИСА ПРОГНОЗИРОВАНИЯ

6.1 Подготовка инструментов

Необходимо установить IDE NetBeans, при установке указать что требуется установить сервер приложений GlassFish. Его можно заменить на аналоги, например Wildfly.

Для начала необходимо создать новый java проект в IDE, и указать что он будет web приложением. Создание такого типа приложения позволит строить распределенную систему из отдельных компонент без привязки к физическому расположению.

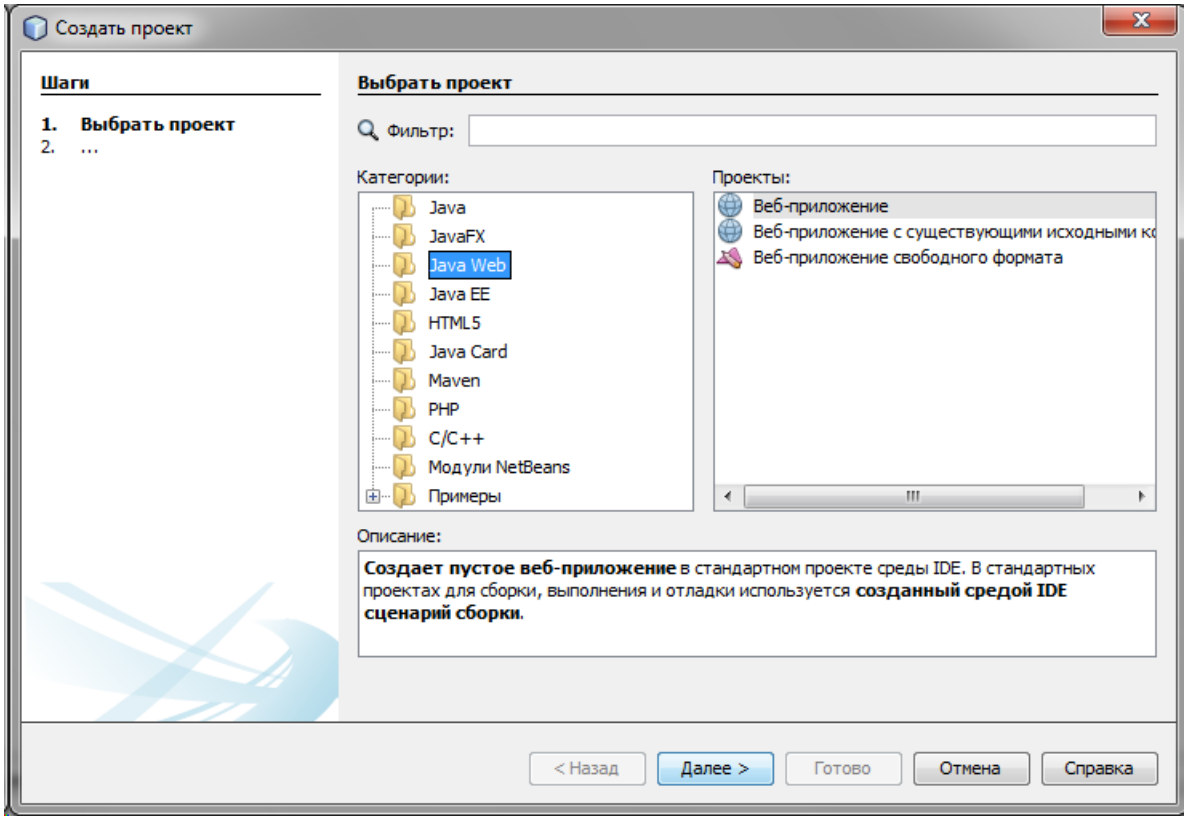

Рис. 6.1. Тип проекта

Задаем наименование проекта, указываем его расположение. Проект, созданный по данной инструкции будет использовать ant в качестве системы управления сборкой.

Далее нужно указать IDE, на какой сервер будет разворачиваться приложение. Эта настройка позволят запускать приложение напрямую из IDE. Разумеется, если приложение впоследствии будет, развернуто на другом сервере то настройки указанные здесь никак не помешают этому.

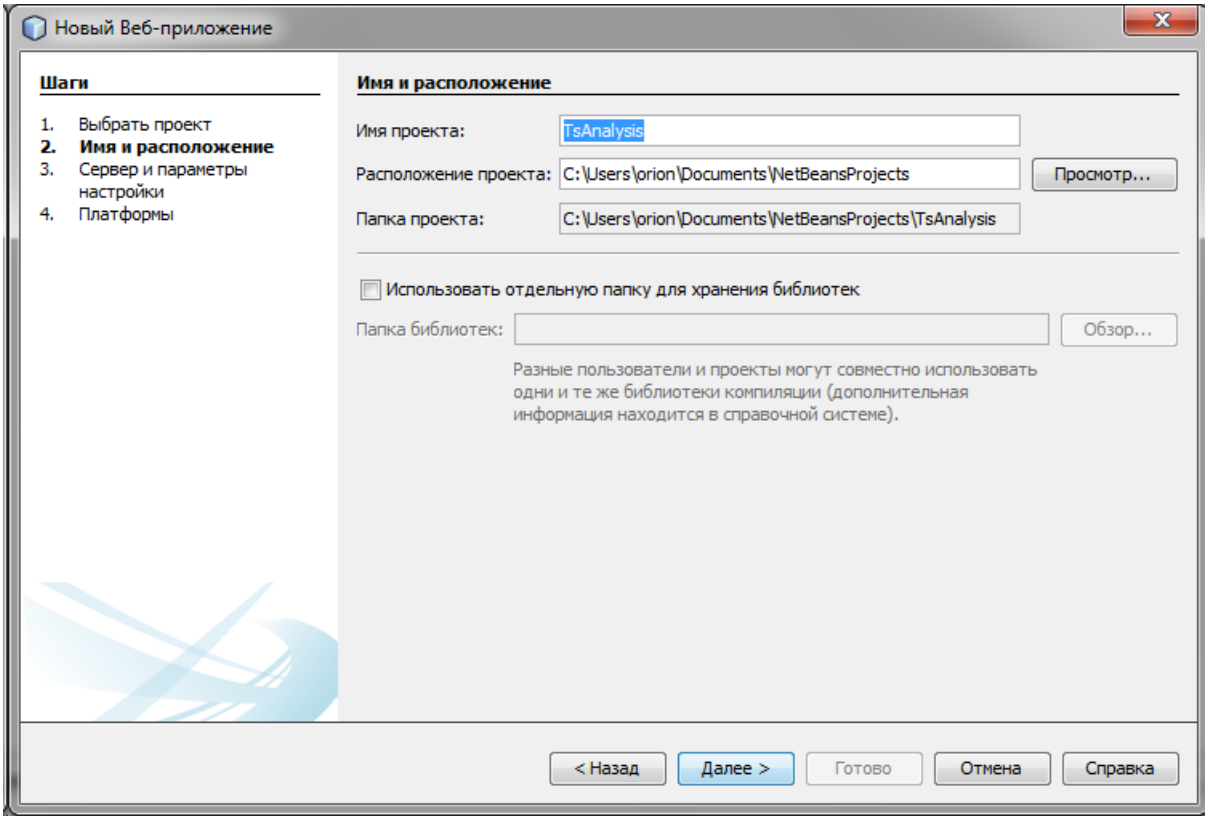

Рис. 6.2. Название проекта

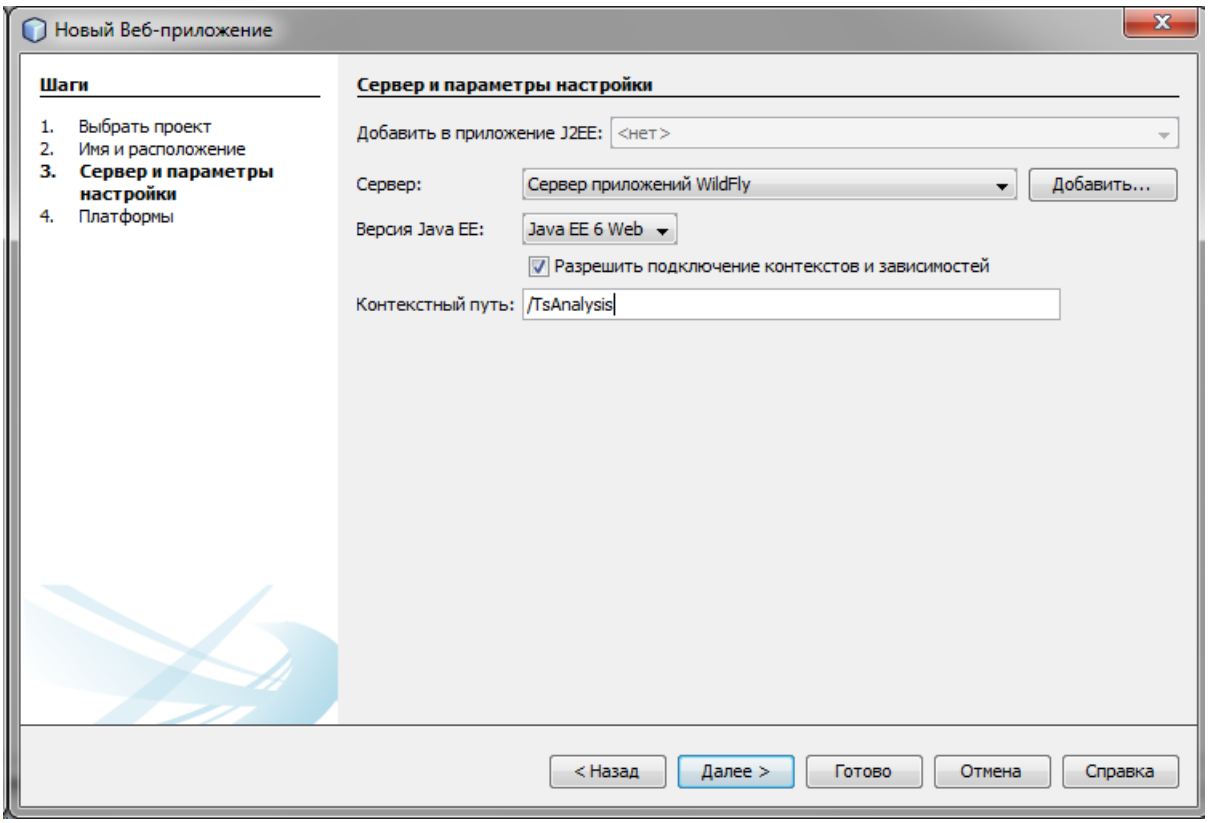

Рис. 6.3. Выбор сервера

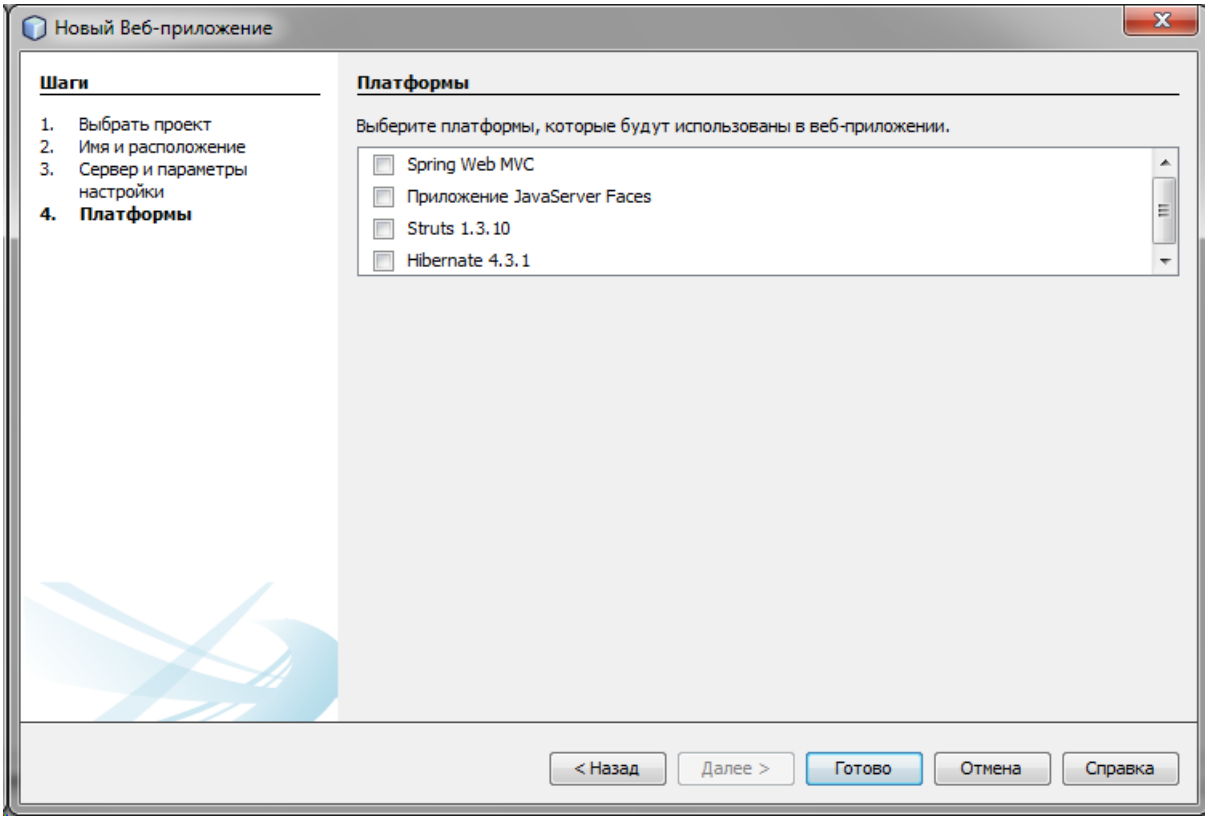

Рис. 6.4. Указание платформ

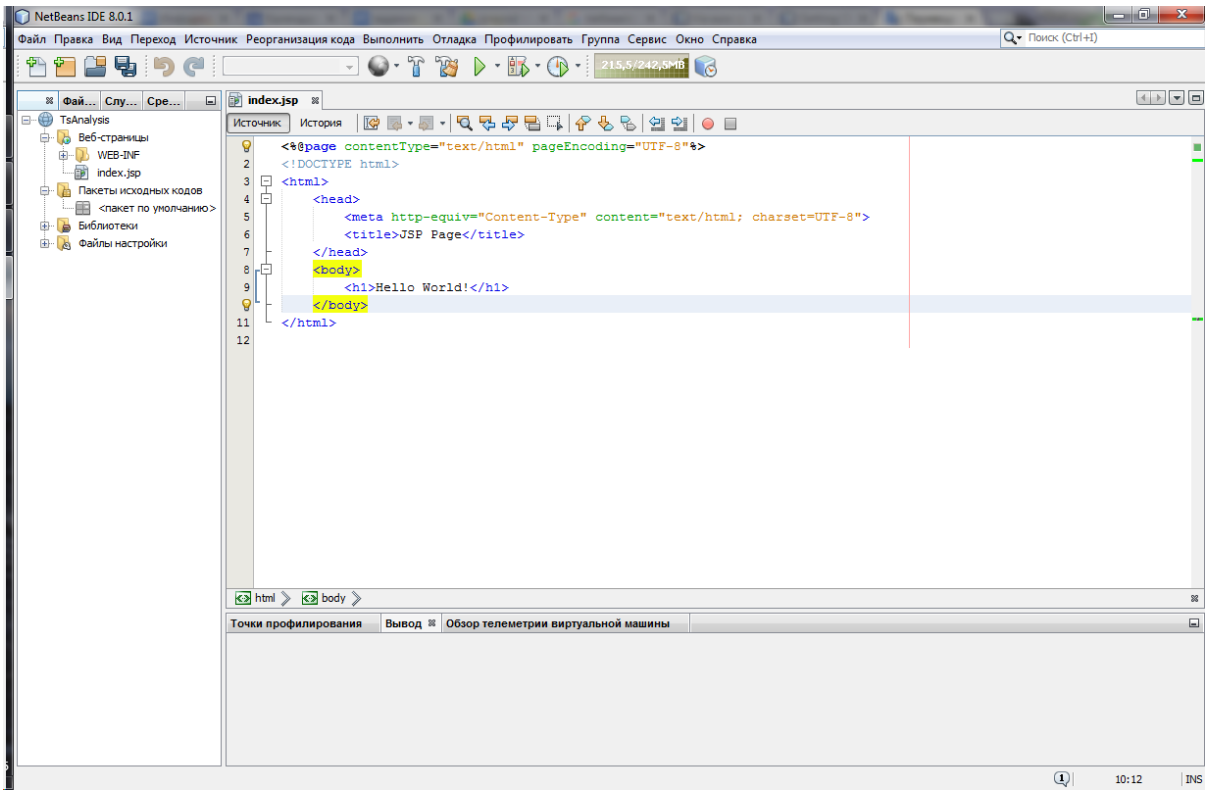

Рис. 6.5. Выбор сервера

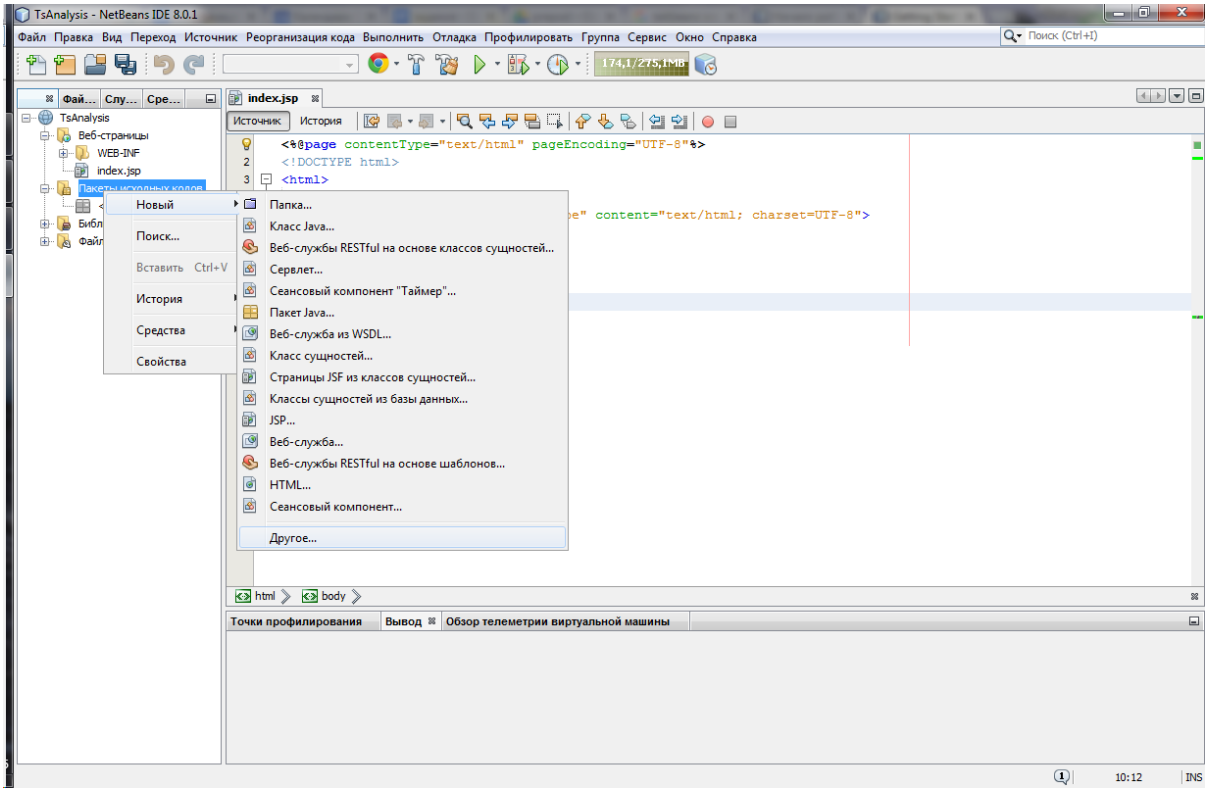

Рис. 6.6. Выбор сервера

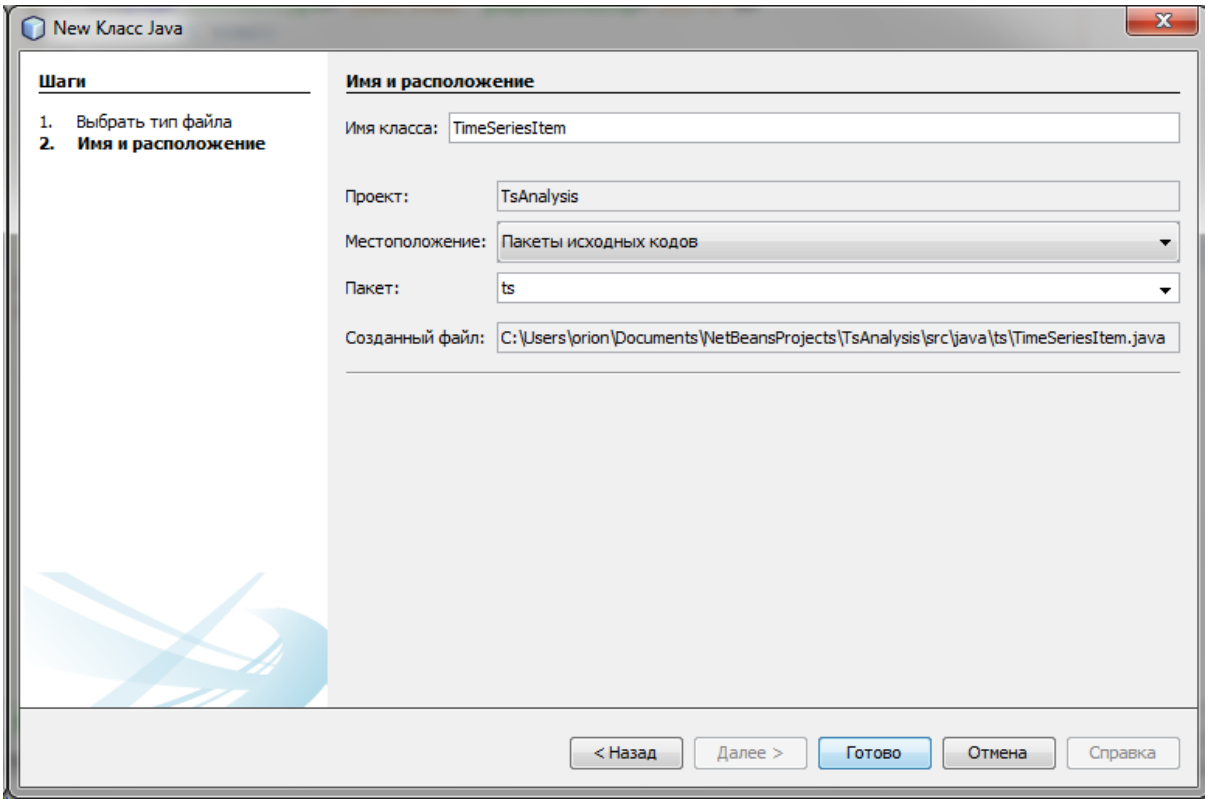

Рис. 6.7. Выбор сервера

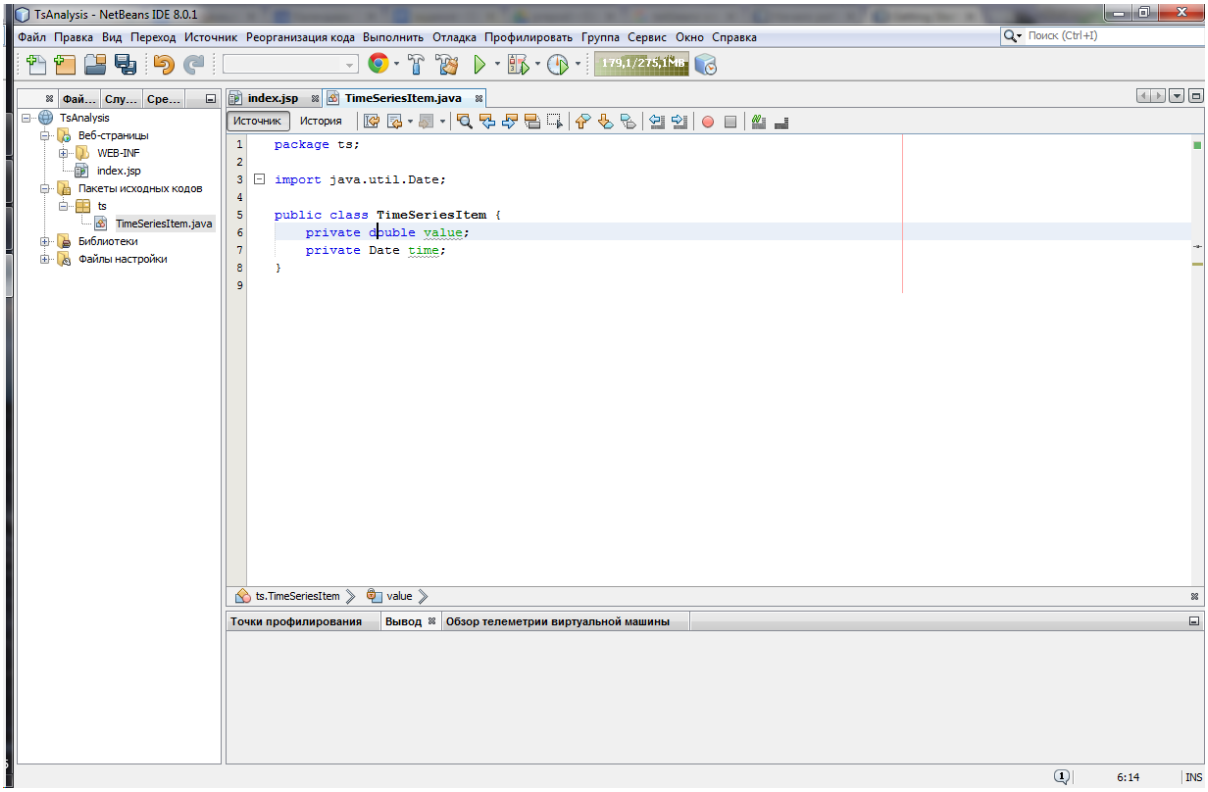

Рис. 6.8. Выбор сервера

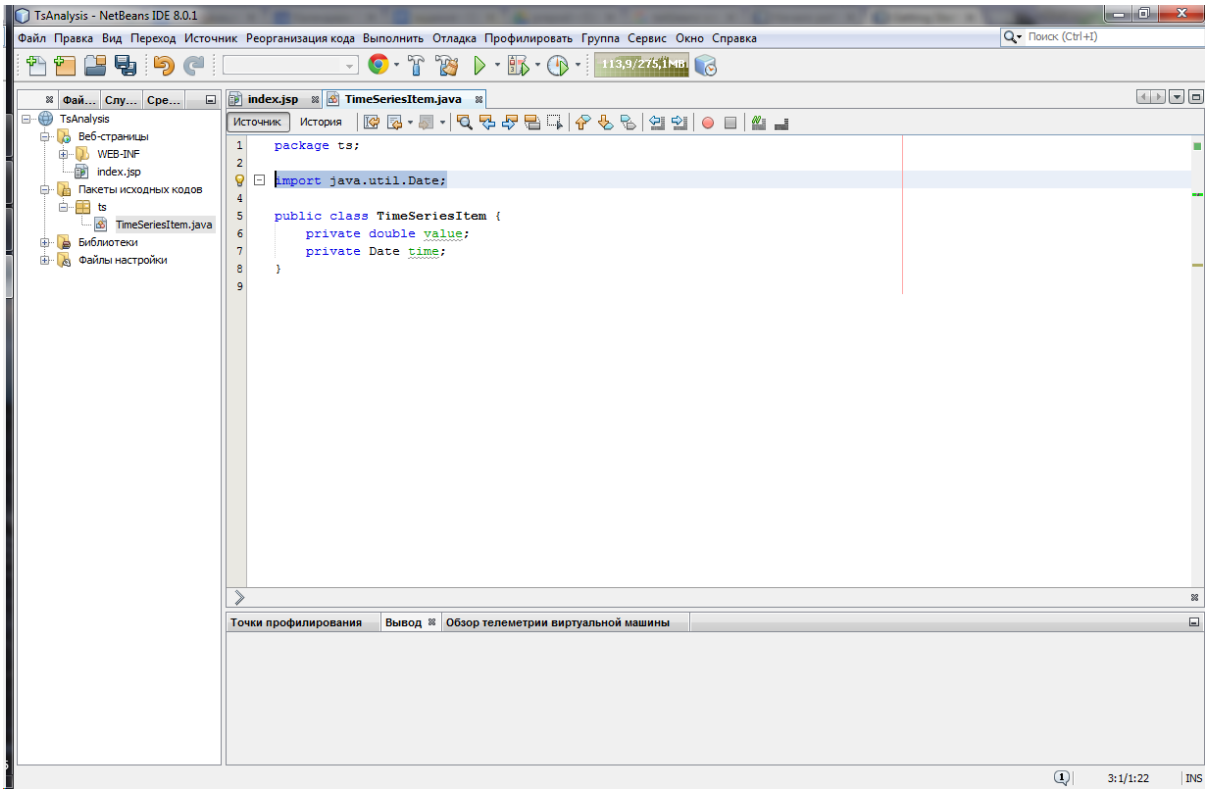

Рис. 6.9. Выбор сервера

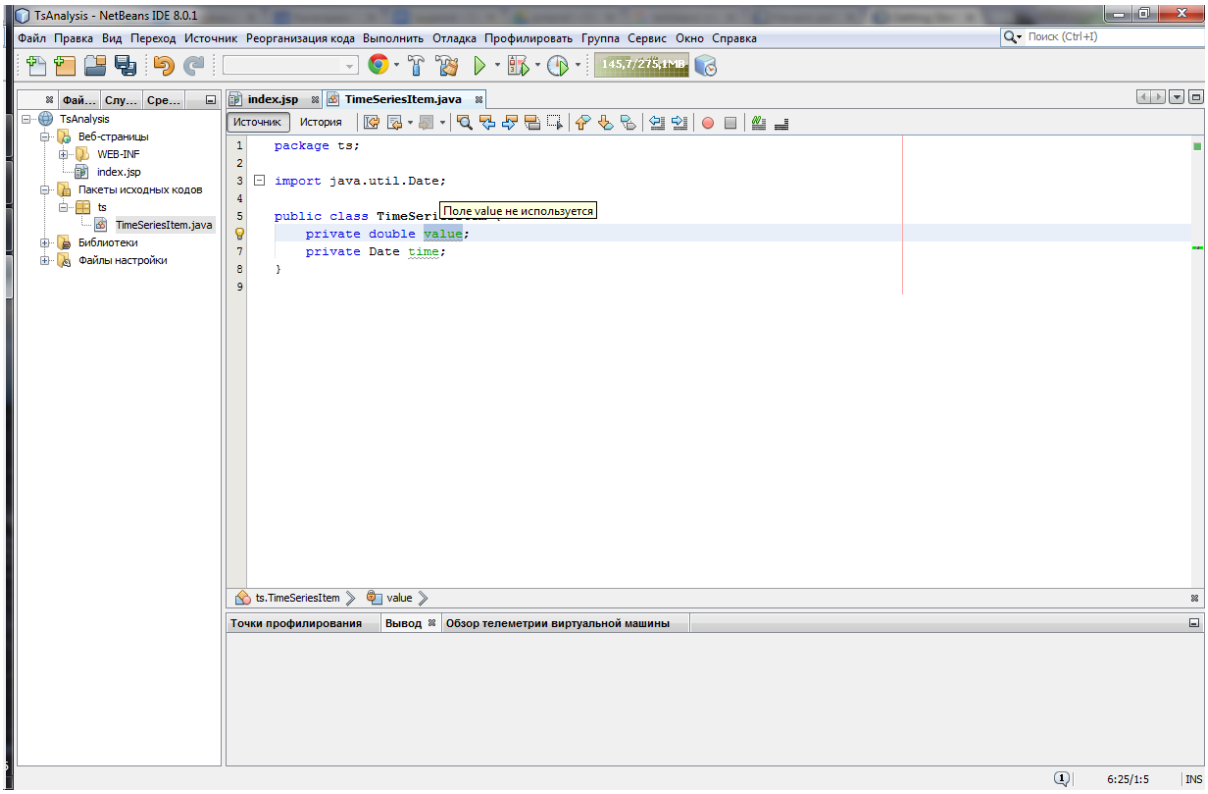

Рис. 6.10. Выбор сервера

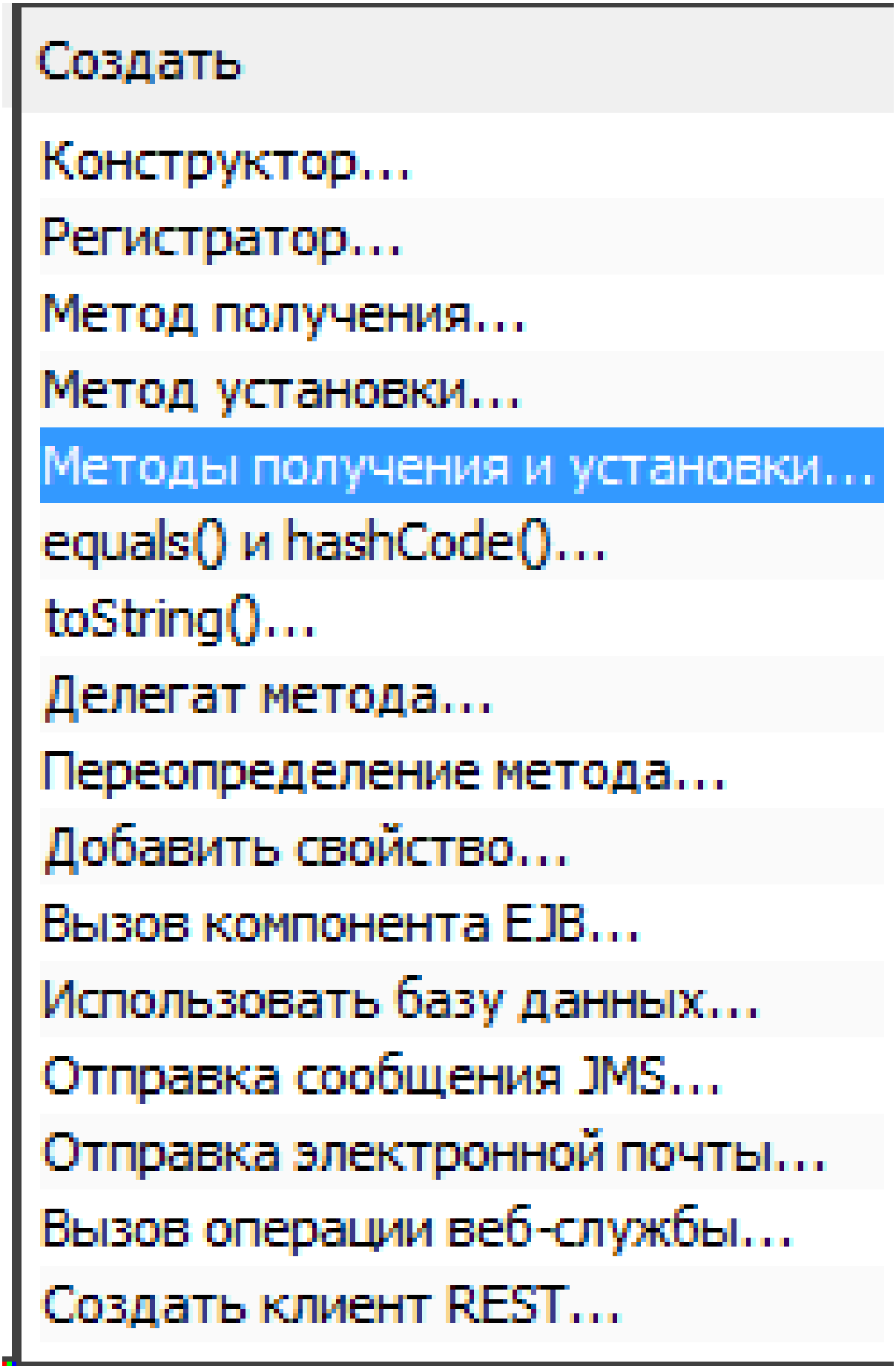

Рис. 6.11. Выбор сервера

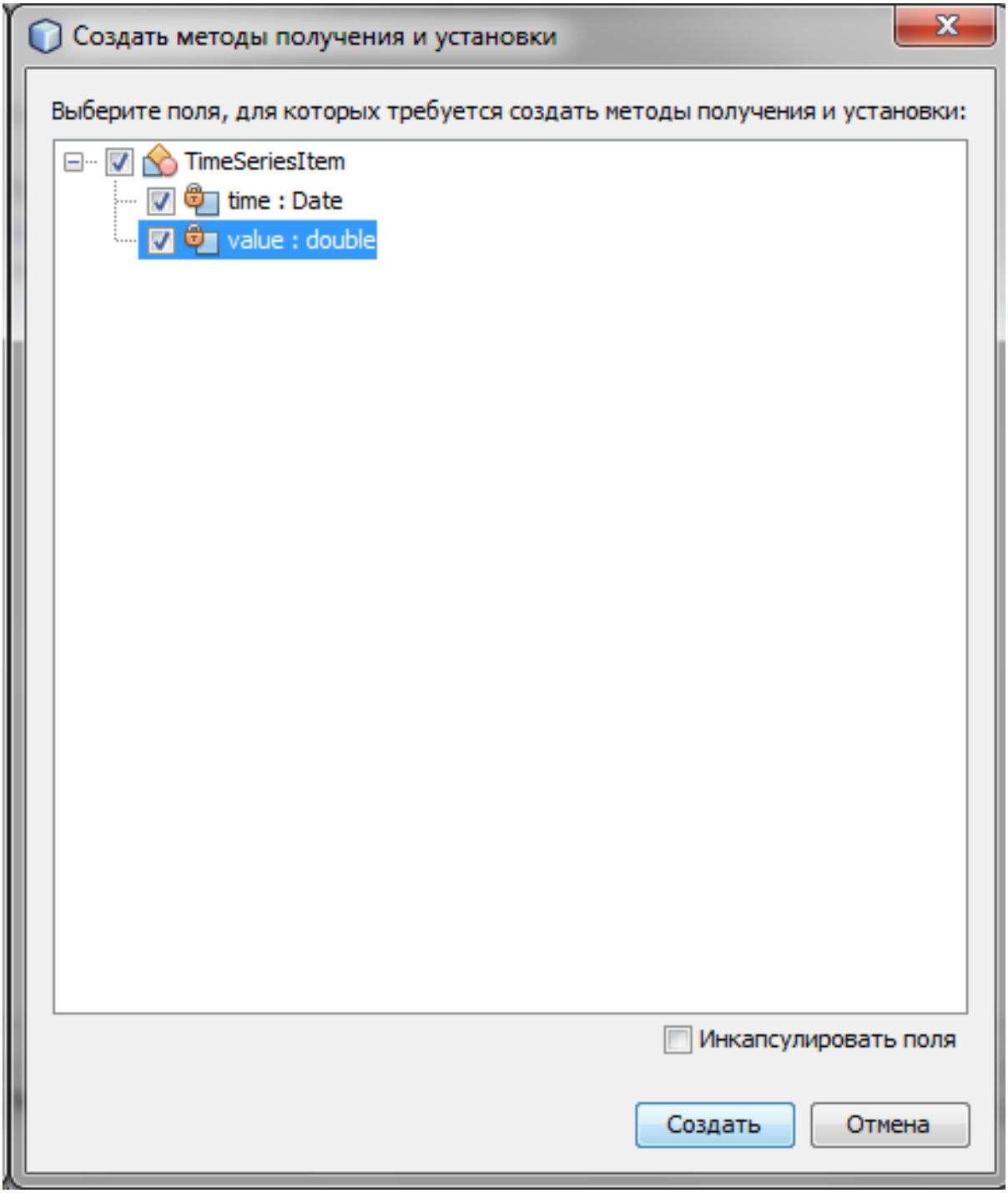

Рис. 6.12. Выбор сервера

ЗАКЛЮЧЕНИЕ

Целью данного пособия является формирование у студентов знаний об аппаратном обеспечении и архитектуре вычислительных систем, принципах построения и функционирования основных устройств данных систем. Представлены основные этапы развития, семейства и типы вычислительных систем.

Также дается пояснение каким образом программы, написанные на языках программирования высокого уровня, выполняются вычислительными системами. Рассмотрены основные уровни архитектуры вычислительных систем:

- цифровой логический уровень;
- уровень микроархитектуры;
- уровень архитектуры набора команд;
- уровень операционной системы;
- уровень ассемблера.

Описаны функциональные особенности и механизмы взаимодействия данных уровней.

Более подробную информацию по вопросам, рассмотренным в данном пособии, можно найти в [1], [2], [3], [4], и [5].

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. Гук М. Аппаратные средства IBM PC. Энциклопедия. – СПб.: Питер, 2004 – 923 с.:ил.
2. Калабеков Б. А. Цифровые устройства и микропроцессорные системы. – М.: Телеком, 2005.
3. Лехин С. Н. Схемотехника ЭВМ. – СПб.: БХВ-Петербург, 2010. – 672 с.
4. Таненбаум Э. Архитектура компьютера – 6-е изд. – СПб.: Питер, 2013. – 816 с.
5. Программирование на ассемблере для начинающих и не только. – URL: http://asmworld.ru/ (дата обращения: 20.12.2014)
6. FASM (flatassembler) – официальный сайт проекта. – URL: http://flatassembler.net/ (дата обращения: 20.12.2014)

Учебное издание

АППАРАТНОЕ ОБЕСПЕЧЕНИЕ ВЫЧИСЛИТЕЛЬНЫХ СИСТЕМ

Учебное пособие

Составители:
МОШКИН Вадим Сергеевич
ФИЛИППОВ Алексей Александрович

Редактор М. В. Теленкова

ЛР № ??? от ??.??.??.
Подписано в печать 25.12.2014. Формат 60x84/16.
Усл. печ. л. 00,00. Тираж 100 экз. Заказ 000.

ИПК «Венец» УлГТУ, 432027, г.Ульяновск, ул. Сев. Венец, д. 32.