# Цели и задачи

Цель:

Разработать программу, для поиска скрытых закономерностей в данных, и интерпретации знаний, необходимых для принятия решений в различных сферах человеческой деятельности.

Задачи:

Изучить предметную область;

Провести анализ и выбрать метод, для поиска закономерностей в данных;

Описать требования к программе;

Разработать программу.

# Актуальность

В связи с совершенствованием технологий записи и хранения данных, налюдей обрушились колоссальные потоки информационных данных в самых различных областях. Деятельность любого предприятия (коммерческого, производственного, медицинского, научного и т.д.) теперь сопровождается регистрацией и записью всех подробностей его деятельности. Многие компании годами накапливают важную бизнес-информацию, надеясь, что она поможет им в принятии решений.

# Задачи прогнозирования

**Задачи прогнозирования** — в прогностике существуют различные частные виды классических задач на прогнозирование. Формулирование таких задач единообразным образом позволяет сравнивать различные методы, предлагаемые различными дисциплинами.

Оценивание кредитоспособности заёмщиков

Эта задача решается банками при выдаче кредитов. Потребность в автоматизации процедуры выдачи кредитов впервые возникла в период бума кредитных карт 60-70-х годов в США и других развитых странах. Объектами в данном случае являются физические или юридические лица, претендующие на получение кредита. В случае физических лиц признаковое описание состоит из анкеты, которую заполняет сам заёмщик, и, возможно, дополнительной информации, которую банк собирает о нём из собственных источников. Примеры бинарных признаков: пол, наличие телефона. Номинальные признаки — место проживания, профессия, работодатель. Порядковые признаки — образование, занимаемая должность. Количественные признаки — сумма кредита, возраст, стаж работы, доход семьи, размер задолженностей в других банках. Обучающая выборка составляется из заёмщиков с известной кредитной историей. В простейшем случае принятие решений сводится к классификации заёмщиков на два класса: «хороших» и «плохих». Кредиты выдаются только заёмщикам первого класса. В более сложном случае оценивается суммарное число баллов (score (англ.)) заёмщика, набранных по совокупности информативных признаков. Чем выше оценка, тем более надёжным считается заёмщик. Отсюда и название — кредитный скоринг. На стадии обучения производится синтез и отбор информативных признаков и определяется, сколько баллов назначать за каждый признак, чтобы риск принимаемых решений был минимален. Следующая задача — решить, на каких условиях выдавать кредит: определить процентную ставку, срок погашения, и прочие параметры кредитного договора. Эта задача также может быть решения методами обучения по прецедентам.

# Knowledge Discovery in Databases

**обнаружение знаний в базах данных**

Knowledge Discovery in Databases (KDD) – это процесс поиска полезных знаний в "сырых" данных. KDD включает в себя вопросы: подготовки данных, выбора информативных признаков, [очистки данных](http://www.basegroup.ru/glossary_ajax/definitions/data_cleaning), применения методов [Data Mining](http://www.basegroup.ru/library/methodology/data_mining/) ([DM](http://www.basegroup.ru/glossary_ajax/definitions/data_mining)), постобработки данных и интерпретации полученных результатов. Безусловно, "сердцем" всего этого процесса являются методы DM, позволяющие обнаруживать знания.

Этими знаниями могут быть правила, описывающие связи между свойствами данных ([деревья решений](http://www.basegroup.ru/glossary_ajax/definitions/decision_trees)), часто встречающиеся шаблоны ([ассоциативные правила](http://www.basegroup.ru/glossary_ajax/definitions/association_rules)), а также результаты [классификации](http://www.basegroup.ru/glossary_ajax/definitions/classification) ([нейронные сети](http://www.basegroup.ru/glossary_ajax/definitions/neuralnet)) и [кластеризации](http://www.basegroup.ru/glossary_ajax/definitions/clustering) данных ([карты Кохонена](http://www.basegroup.ru/glossary_ajax/definitions/som)) и т.д.

Процесс Knowledge Discovery in Databases, состоит из следующих шагов:

1. **Подготовка исходного набора данных.** Этот этап заключается в создании набора данных, в том числе из различных источников, [выбора обучающей](http://www.basegroup.ru/glossary_ajax/definitions/train_set) выборки и т.д. Для этого должны существовать развитые инструменты доступа к различным источникам данных. Желательно иметь поддержку работы с [хранилищами данных](http://www.basegroup.ru/glossary_ajax/definitions/data_warehouse) и наличие семантического слоя, позволяющего использовать для подготовки исходных данных не технические термины, а бизнес понятия.
2. [**Предобработка данных**](http://www.basegroup.ru/glossary_ajax/definitions/preprocessing)**.** Для того чтобы эффективно применять методы Data Mining, следует обратить внимание на вопросы предобработки данных. Данные могут содержать пропуски, шумы, аномальные значения и т.д. Кроме того, данные могут быть избыточны, недостаточны и т.д. В некоторых задачах требуется дополнить данные некоторой априорной информацией. Наивно предполагать, что если подать данные на вход системы в существующем виде, то на выходе получим полезные знания. Данные должны быть качественны и корректны с точки зрения используемого метода DM. Поэтому первый этап KDD заключается в предобработке данных. Более того, иногда размерность исходного пространства может быть очень большой, и тогда желательно применять специальные алгоритмы [понижения размерности](http://www.basegroup.ru/glossary_ajax/definitions/projection). Это как отбор значимых признаков, так и отображение данных в пространство меньшей размерности.
3. **Трансформация, нормализация данных.** Этот шаг необходим для приведения информации к пригодному для последующего [анализа](http://www.basegroup.ru/glossary_ajax/definitions/what_if_analysis) виду. Для чего нужно проделать такие операции, как приведение типов, [квантование](http://www.basegroup.ru/glossary_ajax/definitions/binning), приведение к "скользящему окну" и прочее. Кроме того, некоторые методы анализа, которые требуют, чтобы исходные данные были в каком-то определенном виде. Нейронные сети, скажем, работают только с числовыми данными, причем они должны быть нормализованы.
4. **Data Mining.** На этом шаге применяются различные алгоритмы для нахождения знаний. Это нейронные сети, деревья решений, алгоритмы кластеризации, установления ассоциаций и т.д.
5. **Постобработка данных.** Интерпретация результатов и применение полученных знаний в бизнес приложениях.

# Data Mining

—это обнаружение в исходных данных ранее неизвестных, нетривиальных, практически полезных и доступных интерпретации знаний. Знания должны описывать новые связи между свойствами, предсказывать значения одних признаков на основе других.

Data Mining – это не один, а совокупность большого числа различных методов обнаружения знаний.

**Data Mining**

**Задачи**

**Прогнозирование**

**Оценивание**

**Описательные**

Кластеризация

Поиск ассоциативных правил

**Предсказательные**

Классификация

**Анализ связей (Link Analysis)**

**Визуализация (Visualization, Graph Mining)**

**Обучение с учителем**

**Обучение без учителя**

**Статистика**

**Методы оптимизации**

**Теория баз данных**

**Теория информации**

**Методы**

**Нейросети**

**Деревья решений**

**Генетические алгоритмы**

**Эволюционное программирование**

**Нечеткой логики**

**Ассоциативной памяти**

**Базовые**

**Область применения**

**Розничная торговля**

**Распознование образов**

**Банковское дело**

**Телекоммуникации**

**Страхование**

**Медицина**

**и др**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Метод** | **Точность** | **Масштаби**  **руемость** | **Интерпрети**  **-руемость** | **Пригод-**  **ность к**  **исполь-**  **зованию** | **Трудо-**  **емкость** | **Разносто**  **-**  **ронность** | **Быстрота** | **Популяр-**  **ность,**  **широта**  **исполь-**  **зования** |
| **классические методы (линейная**  **регрессия)** | нейтральная | высокая | высокая /  нейтраль  ная | высокая | нейтральная | нейтральная | высокая | низкая |
| **нейронны**  **е сети** | высокая | низкая | низкая | низкая | нейтральная | низкая | очень низкая | низкая |
| **методы**  **визуали-**  **зации** | высокая | очень низкая | высокая | высокая | очень высокая | низкая | очень низкая | высокая /  нейтраль  ная |
| **деревья**  **решений** | низкая | высокая | высокая | высокая /  нейтраль  ная | высокая | высокая | высокая /  нейтраль  ная | высокая /  нейтраль  ная |
| **полино-**  **миальные**  **нейронны**  **е сети** | высокая | нейтральная | низкая | высокая /  нейтраль  ная | низкая /  нейтральная | нейтральная | низкая /  нейтральная | нейтральная |

Шкала

1 очень низкая

2 низкая

3 Низкая /нейтральная

4 нейтральная

5 высокая /нейтральная

6 Высокая

7 очень Высокая

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Метод** | **Точность** | **Масштаби**  **руемость** | **Интерпрети**  **-руемость** | **Пригод-**  **ность к**  **исполь-**  **зованию** | **Трудо-**  **емкость** | **Разносто**  **-**  **ронность** | **Быстрота** | **Популяр-**  **ность,**  **широта**  **исполь-**  **зования** | **Итог** |
| **классические методы (линейная**  **регрессия)** | 0,571429 | 0,857143 | 0,714286 | 0,857143 | 0,571429 | 0,571429 | 0,857143 | 0,285714 | 0,660714 |
| **нейронны**  **е сети** | 0,857143 | 0,285714 | 0,285714 | 0,285714 | 0,571429 | 0,285714 | 0,142857 | 0,285714 | 0,375 |
| **методы**  **визуали-**  **зации** | 0,857143 | 0,142857 | 0,857143 | 0,857143 | 1 | 0,285714 | 0,142857 | 0,714286 | 0,607143 |
| **деревья**  **решений** | 0,285714 | 0,857143 | 0,857143 | 0,714286 | 0,857143 | 0,857143 | 0,714286 | 0,714286 | **0,732143** |
| **полино-**  **миальные**  **нейронны**  **е сети** | 0,857143 | 0,571429 | 0,285714 | 0,714286 | 0,428571 | 0,571429 | 0,428571 | 0,571429 | 0,553571 |

# Преимущества использования деревьев решений

Рассмотрев основные проблемы, возникающие при построении деревьев, было бы несправедливо не упомянуть об их достоинствах:

* быстрый процесс обучения;
* генерация правил в областях, где эксперту трудно формализовать свои знания;
* извлечение правил на естественном языке;
* интуитивно понятная классификационная модель;
* высокая точность [прогноза](http://www.basegroup.ru/glossary_ajax/definitions/forecust), сопоставимая с другими методами (статистика, [нейронные сети](http://www.basegroup.ru/glossary_ajax/definitions/neuralnet));
* построение непараметрических моделей.

# Построения дерева решений

Пусть нам задано некоторое обучающее множество T, содержащее объекты (примеры), каждый из которых характеризуется m атрибутами (атрибутами), причем один из них указывает на принадлежность объекта к определенному классу.

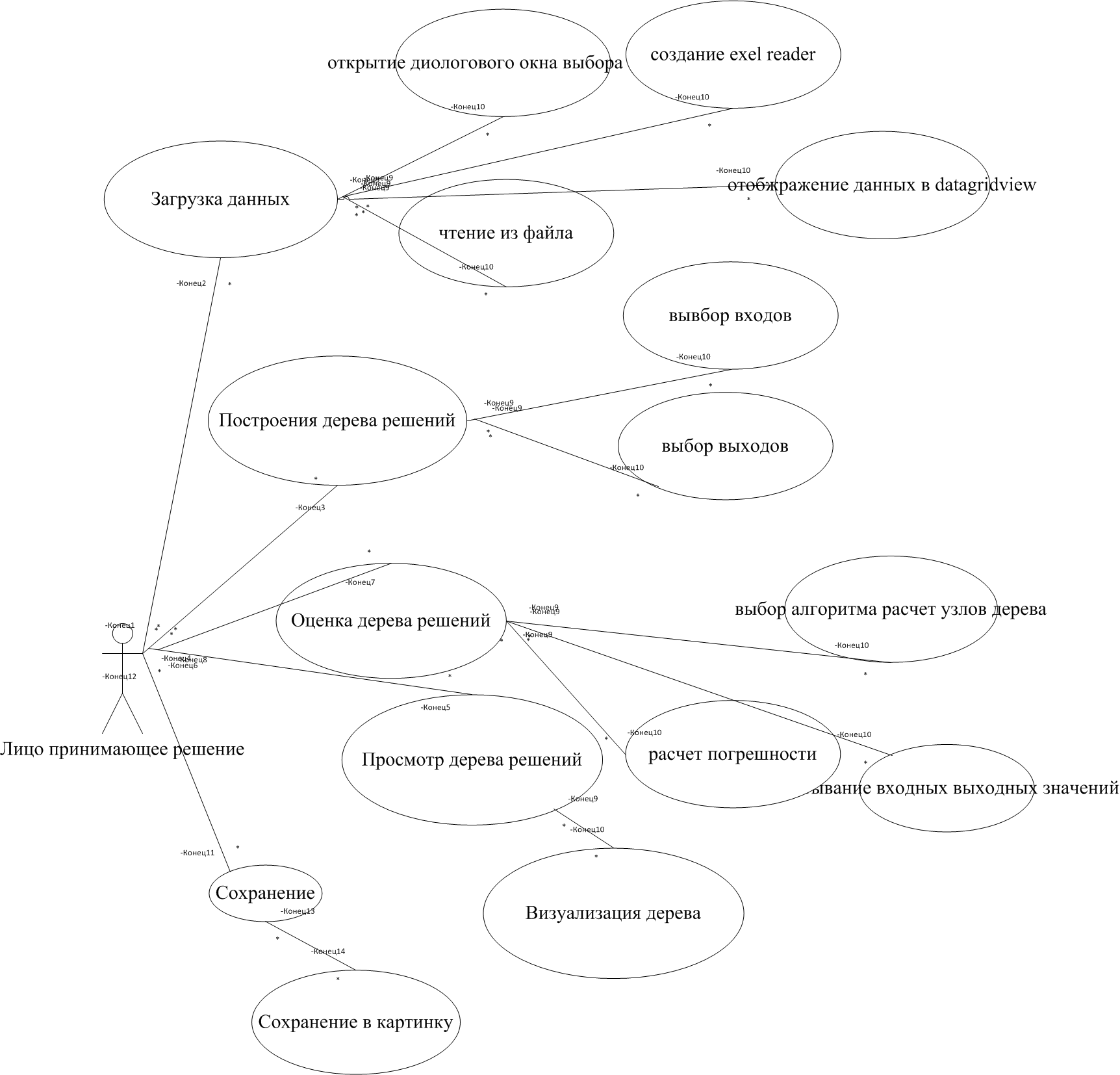
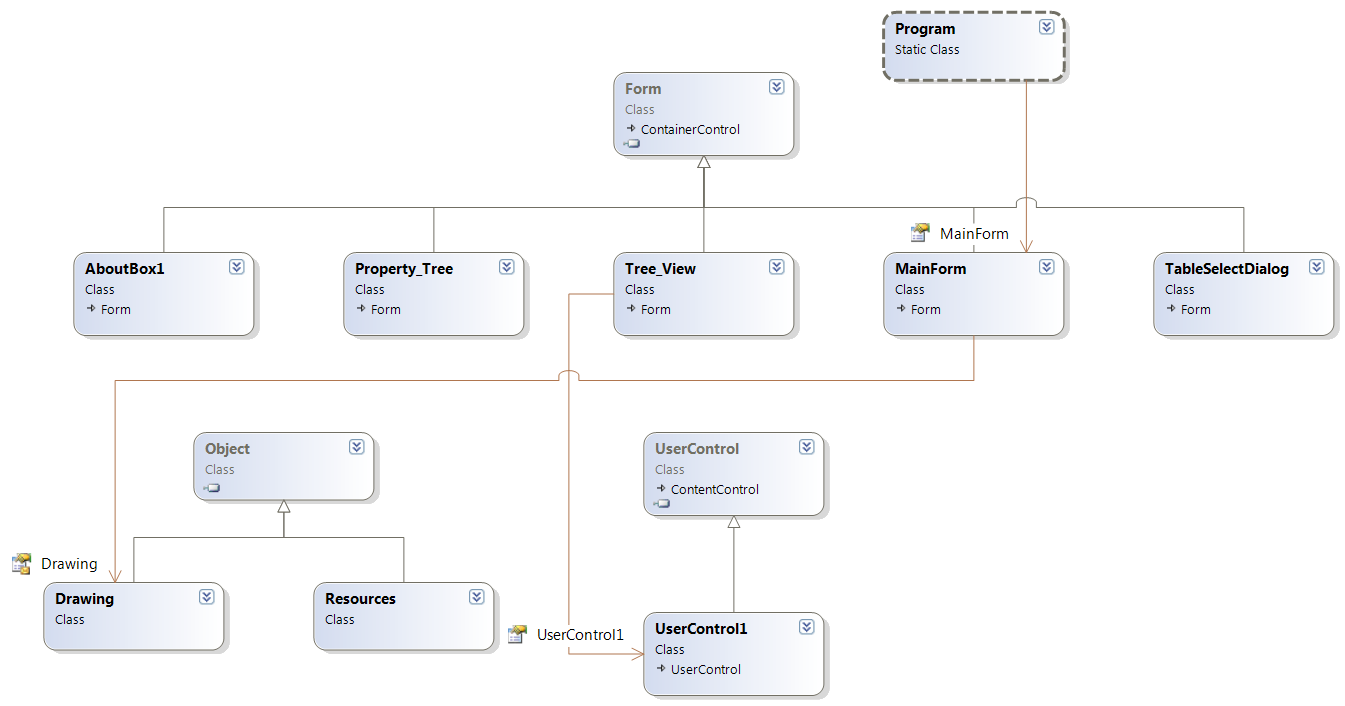
Идею построения деревьев решений из множества T, впервые высказанную Хантом, приведем по Р. Куинлену (R. Quinlan).

Пусть через {C1, C2, ... Ck} обозначены классы(значения метки класса), тогда существуют 3 ситуации:

1. множество T содержит один или более примеров, относящихся к одному классу Ck. Тогда дерево решений для Т – это лист, определяющий класс Ck;
2. множество T не содержит ни одного примера, т.е. пустое множество. Тогда это снова лист, и класс, ассоциированный с листом, выбирается из другого множества отличного от T, скажем, из множества, ассоциированного с родителем;
3. множество T содержит примеры, относящиеся к разным классам. В этом случае следует разбить множество T на некоторые подмножества. Для этого выбирается один из признаков, имеющий два и более отличных друг от друга значений O1, O2, ... On. T разбивается на подмножества T1, T2, ... Tn, где каждое подмножество Ti содержит все примеры, имеющие значение Oi для выбранного признака. Это процедура будет [рекурсивно](http://www.basegroup.ru/glossary_ajax/definitions/recursion) продолжаться до тех пор, пока конечное множество не будет состоять из примеров, относящихся к одному и тому же классу.

# Практическая значимость

Разработки программы для построения модели используя метод дерева решений в задачах прогнозирования

# Выводы:

Проанализирована предметная область;

Проведен анализ и выбран метод, для поиска закономерностей в данных;

Выявлены требования к программе ;

Разработана программа.

*Дальнейшая работа:*

Повысить надежность программы, оптимизировать работоспособность программы;