

МТУСИ

Московский технический университет
связи и информатики

Введение в ИТ

- Anastasia Mozhaeva

Course content

1. Введение. ЭВМ. Практические примеры.

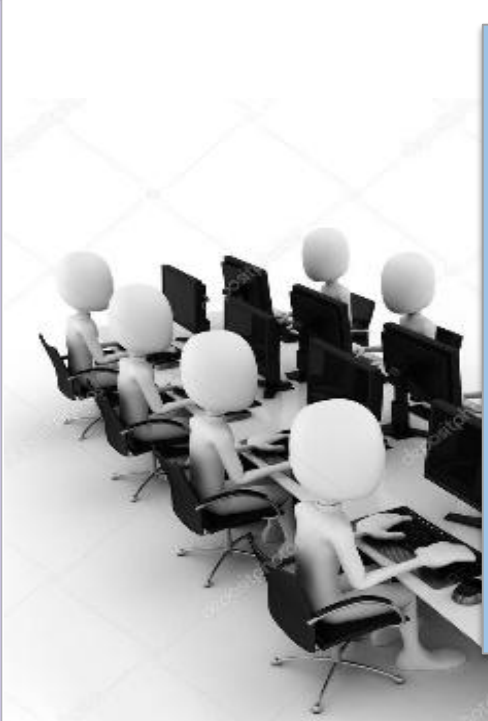
2. Компьютерные комплексы и сети.

3. Программное обеспечение и Интеллектуальный анализ данных.

3.1. Основы интеллектуального анализа

3.2. Сбор данных для интеллектуального анализа

3.3. Алгоритмы машинного обучения и модель интеллектуального анализа данных



Информация – это сведения о лицах, предметах, фактах, событиях, явлениях, процессах независимо от формы их представления.

Три аспекта информации:
Прагматический аспект
Семантический аспект
Синтаксический аспект

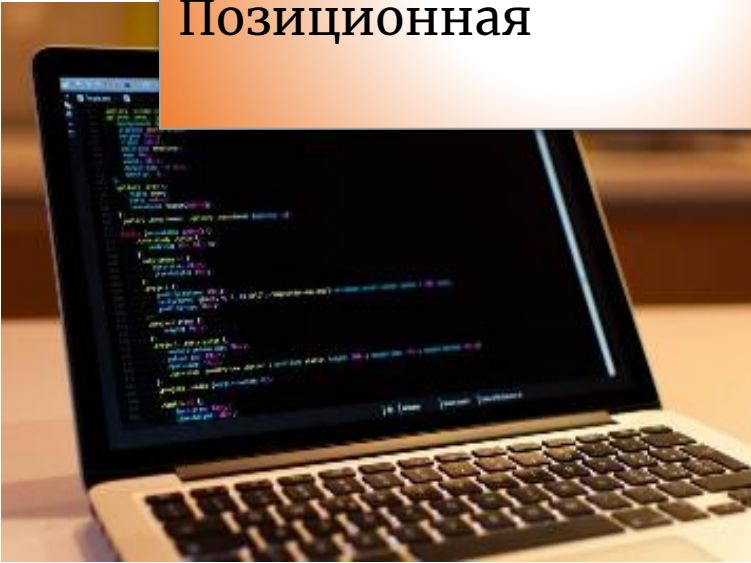
Информационная технология (ИТ) – процесс, использующий совокупность средств и методов сбора, обработки и передачи первичной информации для получения информации нового качества о состоянии объекта, процесса или явления (информационного продукта).

Три уровня ИТ:
– теоретический
– исследовательский
– прикладной



Представление данных.

Системы счисления:
Не позиционная
Позиционная

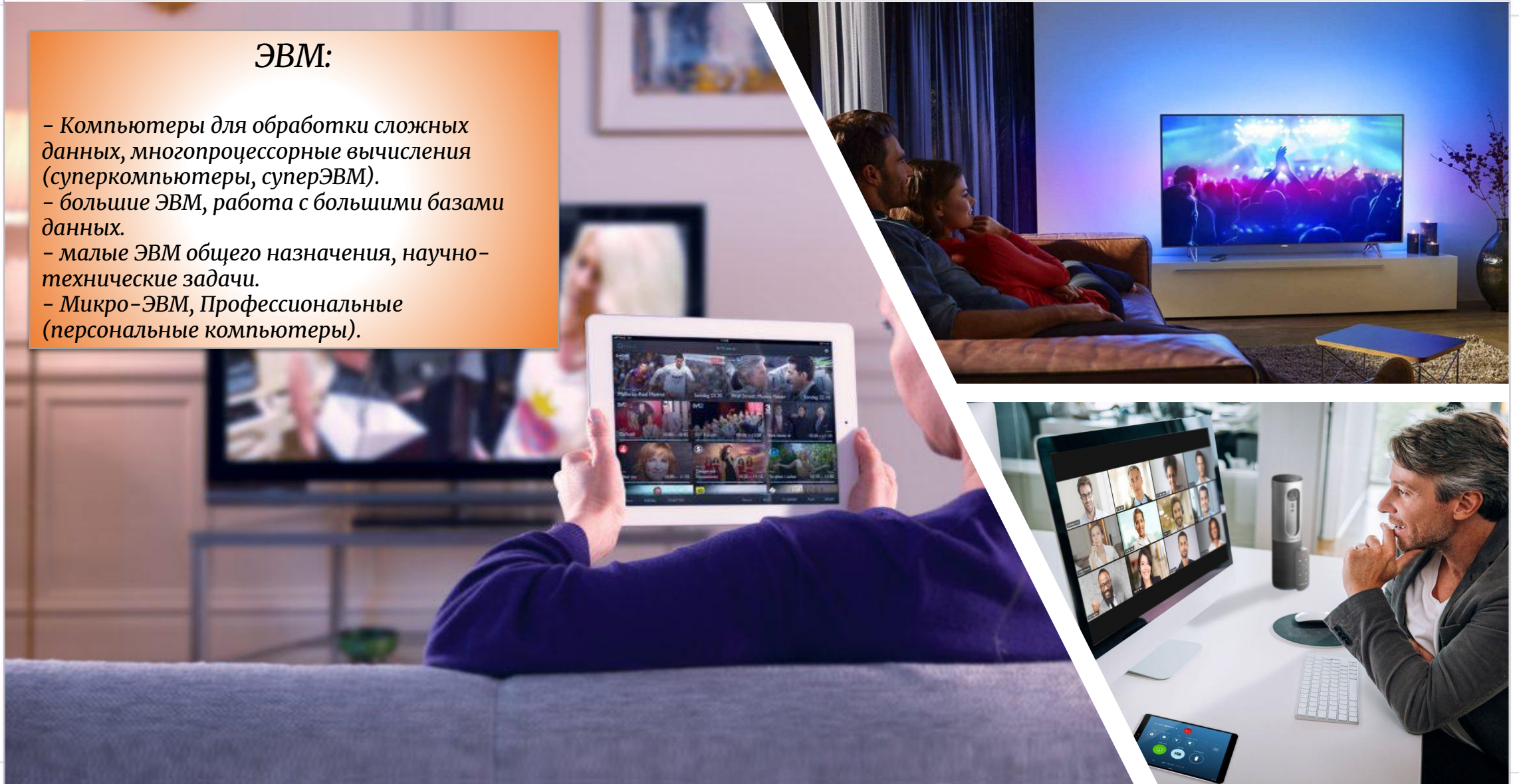


Перевод числа из десятичной системы в двоичную

$$\begin{array}{r}
 100 \quad | \quad 2 \\
 \hline
 100 \quad 50 \\
 \hline
 \mathbf{0}
 \end{array}
 \quad
 \begin{array}{r}
 50 \quad | \quad 2 \\
 \hline
 50 \quad 25 \\
 \hline
 \mathbf{0}
 \end{array}
 \quad
 \begin{array}{r}
 25 \quad | \quad 2 \\
 \hline
 24 \quad 12 \\
 \hline
 \mathbf{1}
 \end{array}
 \quad
 \begin{array}{r}
 12 \quad | \quad 2 \\
 \hline
 12 \quad 6 \\
 \hline
 \mathbf{0}
 \end{array}
 \quad
 \begin{array}{r}
 6 \quad | \quad 2 \\
 \hline
 6 \quad 3 \\
 \hline
 \mathbf{0}
 \end{array}
 \quad
 \begin{array}{r}
 3 \quad | \quad 2 \\
 \hline
 2 \quad 1 \\
 \hline
 \mathbf{1}
 \end{array}$$

ЭВМ:

- Компьютеры для обработки сложных данных, многопроцессорные вычисления (суперкомпьютеры, суперЭВМ).
- большие ЭВМ, работа с большими базами данных.
- малые ЭВМ общего назначения, научно-технические задачи.
- Микро-ЭВМ, Профессиональные (персональные компьютеры).



Пять базовых элементов компьютера, согласно Джон фон Неймана:

- арифметико-логическое устройство (арифметические и логические операции над данными);*
- устройство управления (управление аппаратными и программными ресурсами);*
- запоминающее устройство;*
- система ввода информации;*
- система вывода информации.*

Программное
обеспечение
(ПО) –
организованная
совокупность
обрабатывающих
программ и
обрабатываемых
данных

Общее ПО – предназначено для обеспечения функционирования компьютера и эффективной работы на нём. Этим ПО пользуется каждый пользователь. В состав ПО входит: операционная система (ОС) и специальный комплекс программ технического обслуживания (КПТО).

Специальное (или прикладное) ПО – предназначено для решения специальных прикладных задач. С ним работают пользователи-специалисты какой либо прикладной области

Системы программирования

Системы программирования предназначены для автоматизации процесса написания программ. В их состав входит язык программирования (ЯП), транслятор (Т) и специальные средства редактирования, отладки и компоновки (СРОК).

Язык программирования – совокупность правил, определяющих систему записей, составляющих программу, а так же определяющих синтаксис и семантику (смысл) используемых грамматических конструкций.

Вычислительные комплексы и сети

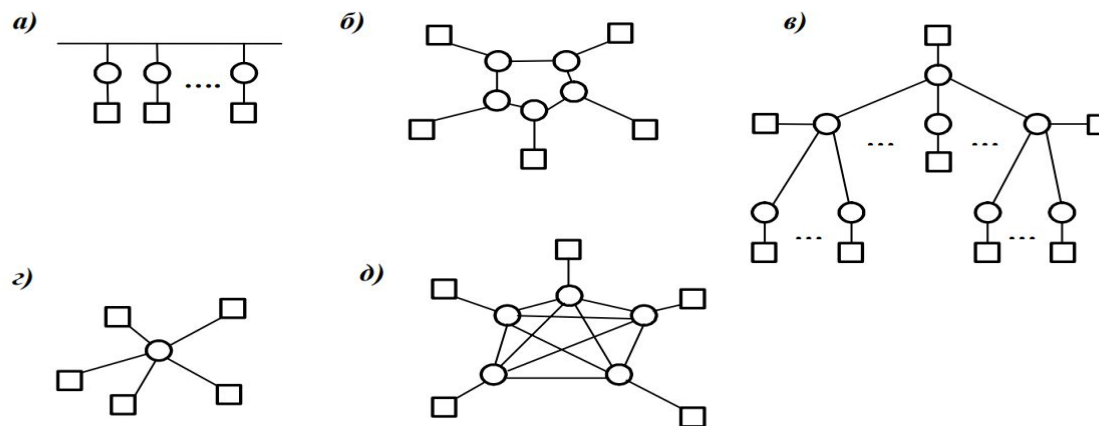
Компьютерная сеть представляет собой совокупность компьютеров, объединенных средствами передачи данных.

Архитектура сети ЭВМ определяет принципы построения и функционирования аппаратного и программного обеспечения элементов сети.

Обработка информации при помощи ЭВМ развивается по двум направлениям:

- с использованием вычислительных комплексов;
- с использованием вычислительных сетей.

Вычислительные комплексы объединяют несколько ЭВМ, территориально расположенных в одном месте.



Типы структур компьютерных сетей: а) - общая шина; б) - кольцо; в) - иерархическая структура; г) - радиальная (звезда); д) - многозвенная;

Передача нескольких видеопотоков по одному каналу связи: уменьшение объема передаваемых данных на два порядка.

60-minute = 670 GB

Бытовая проводная сеть передает около 360 ГБ в час.

Передача данных по WIFI, это на порядок медленнее, чем в проводной сети.

1. Пример

Сжатие без потерь

Может восстановить всю исходную информацию из сжатых данных

Сжатие с потерями

Гораздо большее сжатие за счет уменьшения информации. Не принципиальная, избыточная информация для восприятия зрительной системой человека удаляется или сокращается, а это влияет на качество.

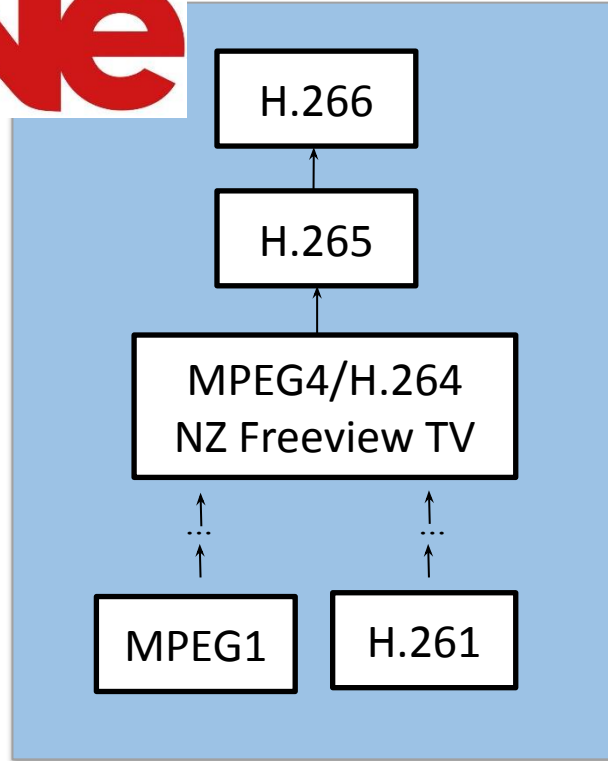


Lossy Compression

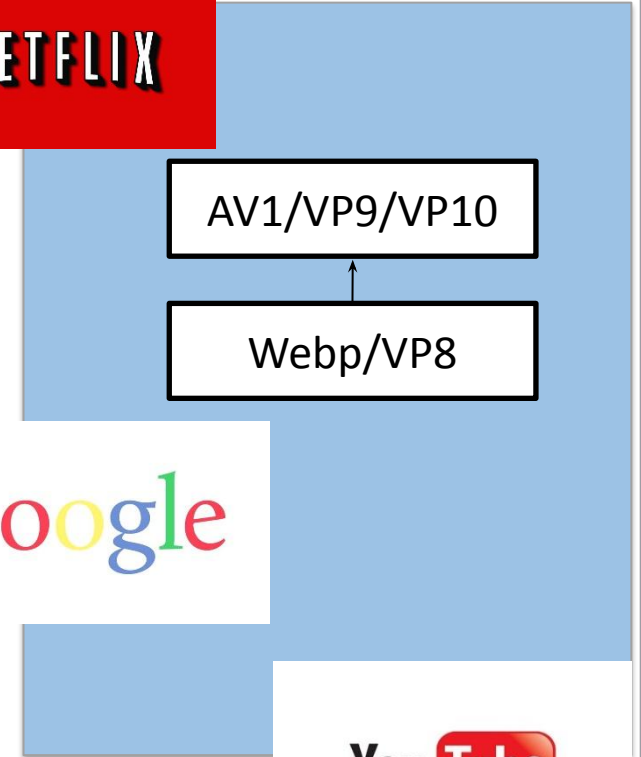


1.1. Modern video lossy compression methods

ONE



NETFLIX



Google

You Tube

The neural network compression

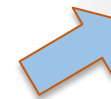
Проблемы кроются в работе алгоритмов адаптации и скорости передачи данных.

1.2. Quality assessment

Test



Субъективное качество



Reference



Объективное качество



Оценка качества - это характеристика обработанного видео по сравнению с оригиналом.

1.3. The current models used by quality assessment

Peak signal-to-noise ratio (PSNR)
Structural similarity image metric (SSIM)

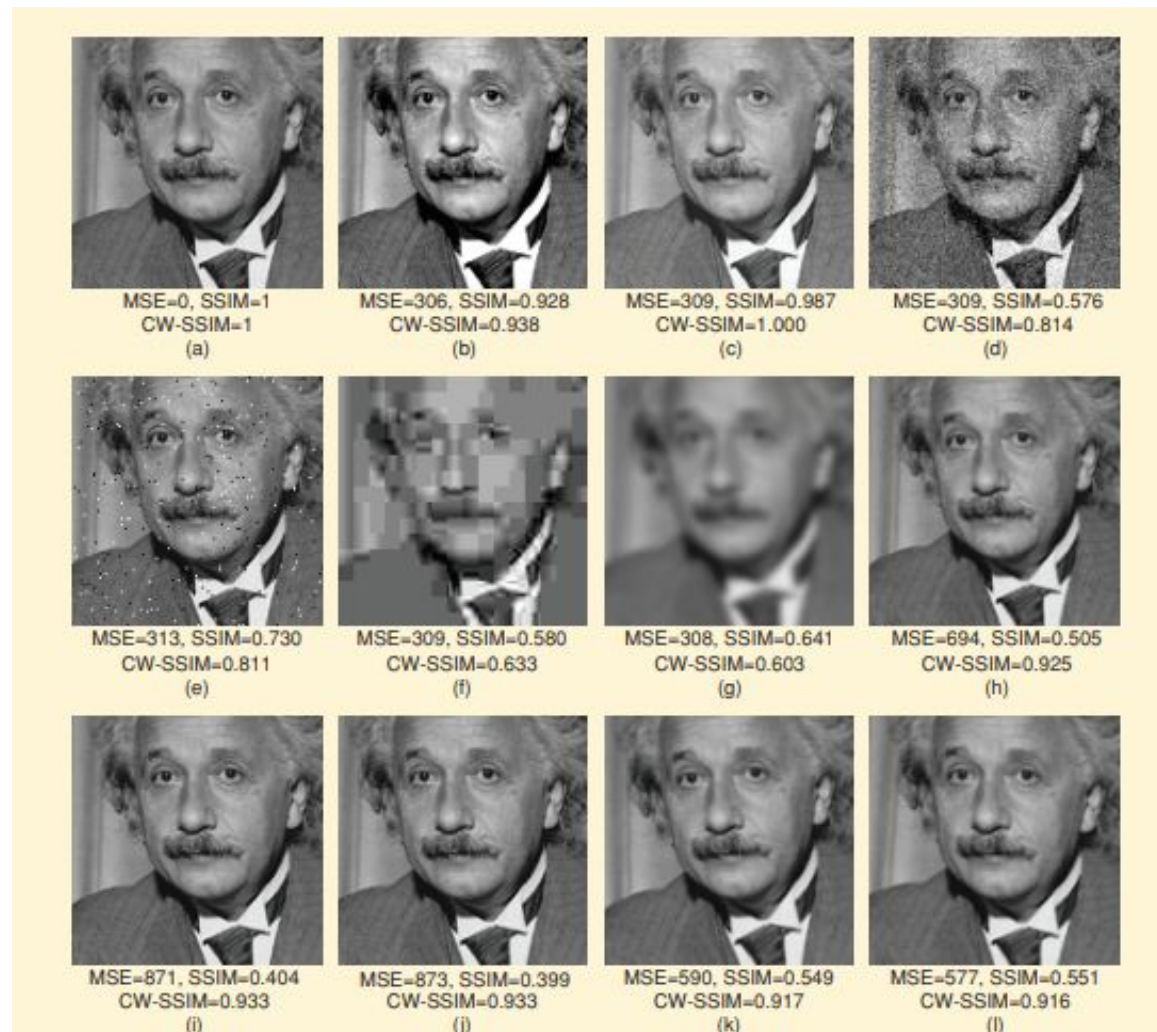
Преимущества

Вычислить это просто и недорого. Это имеет ясный физический смысл. Отличная метрика в контексте оптимизации. Широко используется просто потому, что это соглашение.

Недостатки

Меры неверно отражают структурные перекосы. Плохо коррелируют с визуальной оценкой качества. Местные оценки SSIM нестабильны. Не учитывайте разные абсолютные уровни яркости или расстояние просмотра.

1.4. The current models used by quality assessment



Comparison of image fidelity measures for “Einstein” image altered with different types of distortions, / Zhou Wang , Alan C. Bovik , Ligang Lu

1.5. Возможные решения

Создание новых алгоритмов качества, использующих языки программирования

Создание новых баз субъективного качества, использующих интеллектуальный анализ данных

Weka

Интеллектуальный анализ данных с помощью Weka

Объяснение принципов популярных алгоритмов

Практика

Опыт в области фактического анализа данных

2. Интеллектуаль ный анализ данных. Weka.

Интеллектуальный анализ данных - это переход от необработанных данных к информации, которая может использоваться для предсказаний, полезных в реальном мире.

1. **Сбор данных – это приложение**
2. **Машинное обучение – это алгоритмы**

2. Интеллектуаль ный анализ данных. Weka.

Идеальная ситуация

- 1:** У нас много исторических данных
- 2:** у нас есть данные о текущей ситуации
- 3:** и мы хотим выбрать лучший вариант

2. Интеллектуаль ный анализ данных. Weka.

RQ: «Что такое Weka?»

- Птичка?
- Среда для анализа знаний?

2. Интеллектуаль ный анализ данных. Weka.

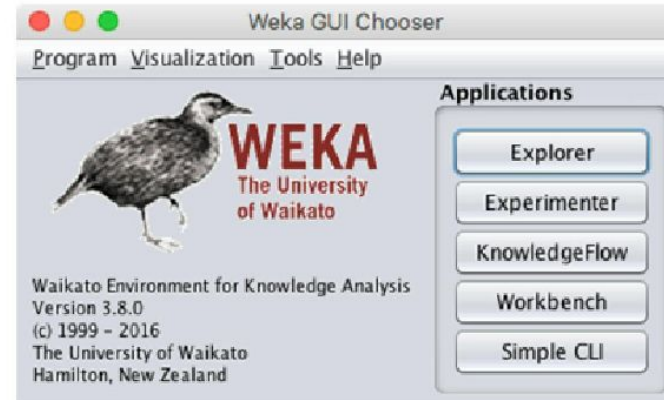
Установка Weka: предварительный просмотр

<http://www.cs.waikato.ac.nz/ml/weka.>

- Нажмите кнопку *Загрузить и установить*
- Выберите, подходящую версию для вашего компьютера; Windows, Mac OS или Linux
- После загрузки, открывайте загрузку. Просто продолжайте нажимать «Далее»! Установите его на место по умолчанию - и запомните название этого места!
- Можете создать ярлык и поместить его на рабочий стол для удобства.
- Сделайте копию папки *данные* (в папке Weka) и поместите ее в удобное место для дальнейшего использования

2. Интеллектуаль ный анализ данных. Weka.

Установка Weka



Сравнение
производительности

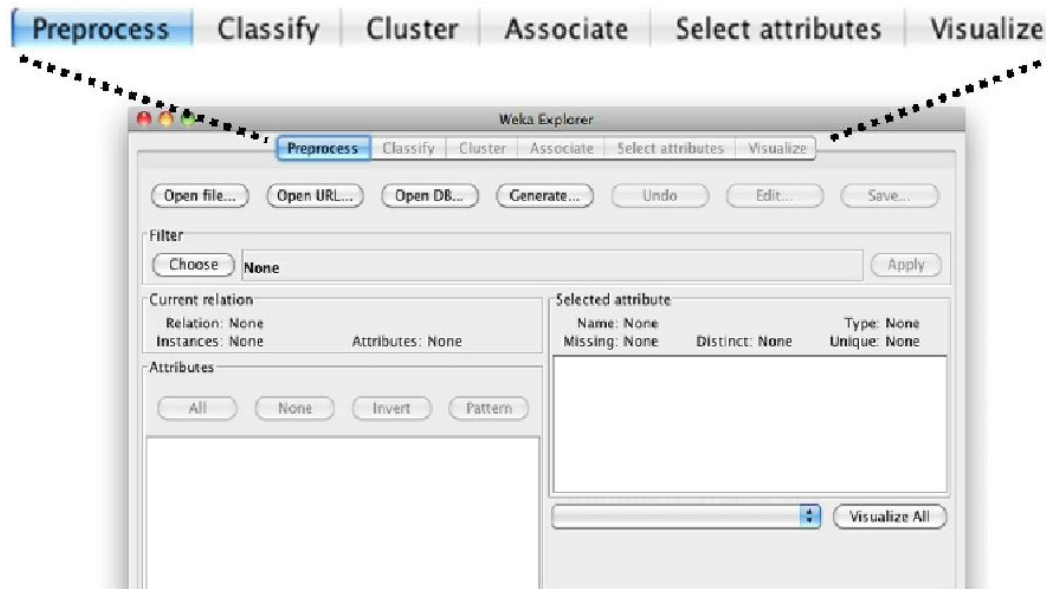
Графический
интерфейс

Общий
интерфейс

Интерфейс
командной строки

2. Интеллектуаль ный анализ данных. Weka.

Установка Weka



2. Интеллектуаль ный анализ данных. Weka.

Интеллектуальный анализ данных с помощью Weka

Набор данных - это набор экземпляров.

Экземпляр - это единственный пример.

Атрибут - это характеристика экземпляра.

Цель - определить класс новых экземпляров.

Классификатор - это модель, подобная некоторой формуле, которая позволяет определять атрибут класса из других атрибутов.

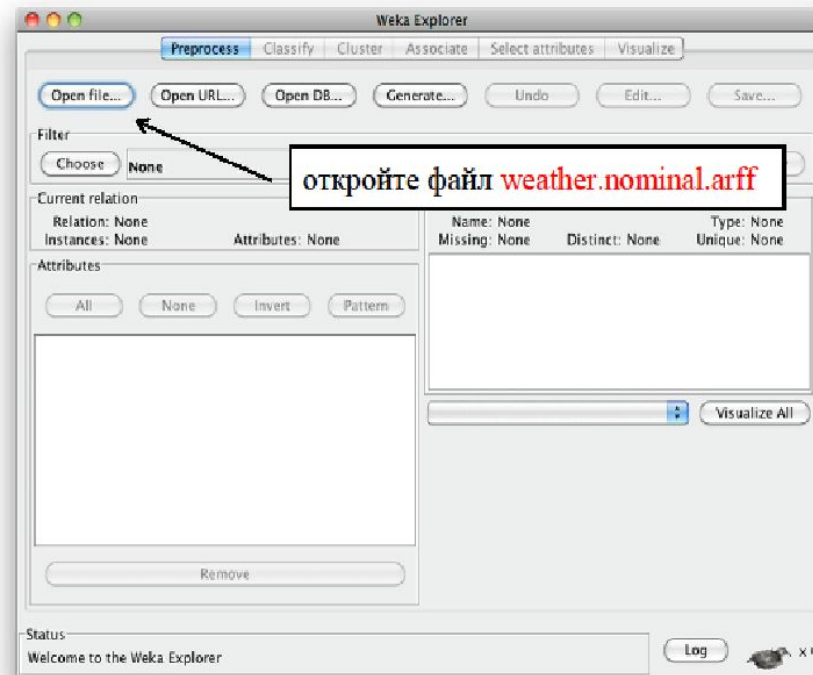
2. Интеллектуальн ый анализ данных. Weka.

Набор данных Weather

		атрибуты				
		Outlook	Temp	Humidity	Windy	Play
экземпляры	1	Sunny	Hot	High	False	No
	2	Sunny	Hot	High	True	No
	3	Overcast	Hot	High	False	Yes
	4	Rainy	Mild	High	False	Yes
	5	Rainy	Cool	Normal	False	Yes
	6	Rainy	Cool	Normal	True	No
	7	Overcast	Cool	Normal	True	Yes
	8	Sunny	Mild	High	False	No
	9	Sunny	Cool	Normal	False	Yes
	10	Rainy	Mild	Normal	False	Yes
	11	Sunny	Mild	Normal	True	Yes
	12	Overcast	Mild	High	True	Yes
	13	Overcast	Hot	Normal	False	Yes
	14	Rainy	Mild	High	True	No

2. Интеллектуаль ный анализ данных. Weka.

Набор данных Weather



2. Интеллектуальн ый анализ данных. Weka.

Набор данных Weather

атрибуты

значение атрибутов

No.	Label	Count
1	sunny	5
2	overcast	4
3	rainy	5

Class	Count
play (No)	5
play (Yes)	5

2. Интеллектуальн ый анализ данных. Weka.

Еще данные о погоде

		атрибуты			
		Outlook	Temp	Humidity	Windy
экземпляры	1	Sunny	Hot	High	False
	2	Sunny	Hot	High	True
	3	Overcast	Hot	High	False
	4	Rainy	Hot	High	False
	5	Rainy	Cool	Normal	False
	6	Rainy	Cool	Normal	True
	7	Rainy	Cool	Normal	True
	8	Sunny	Cool	Normal	True
	9	Sunny	Cool	Normal	False
	10	Rainy	Mild	Normal	False
	11	Sunny	Mild	Normal	True
	12	Overcast	Mild	High	True
	13	Overcast	Hot	Normal	False
	14	Rainy	Mild	High	True

Задача классификации:
определить значение "класса"

2. Интеллектуальный анализ данных. Weka.

Еще данные о погоде

откройте файл `weather.nominal.arff`

атрибуты

класс

значение атрибутов

No.	Label	Count
1	sunny	5
2	overcast	4
3	rainy	5

No.	Name
1	outlook
2	temperature
3	humidity
4	windy
5	play

Class: play (Nom) Visualize All

2. Интеллектуальный анализ данных. Weka.

Еще данные о погоде

Классификация

иногда называется «контролируемое обучение»

Набор данных: классифицированные примеры



“ Модель”, которая классифицирует новые примеры



2. Интеллектуальный анализ данных. Weka.

Еще данные о погоде

откройте файл **weather.numeric.arff**

атрибуты

класс

No.	Label	Count
1	sunny	5
2	overcast	4
3	rainy	5

значение атрибутов

2. Интеллектуальн ый анализ данных. Weka.

Интеллектуальный анализ данных с помощью Weka

```
@relation weather
```

```
@attribute outlook {sunny, overcast, rainy}
```

```
@attribute temperature numeric
```

```
@attribute humidity numeric
```

```
@attribute windy {TRUE, FALSE}
```

```
@attribute play {yes, no}
```

```
@data
```

```
sunny,85,85,FALSE,no
```

```
sunny,80,90,TRUE,no
```

```
overcast,83,86,FALSE,yes
```

```
rainy,70,96,FALSE,yes
```

```
rainy,68,80,FALSE,yes
```

```
rainy,65,70,TRUE,no
```

```
overcast,64,65,TRUE,yes
```

```
sunny,72,95,FALSE,no
```

```
sunny,69,70,FALSE,yes
```

```
rainy,75,80,FALSE,yes
```

```
sunny,75,70,TRUE,yes
```

```
overcast,72,90,TRUE,yes
```

```
overcast,81,75,FALSE,yes
```

```
rainy,71,91,TRUE,no
```


2. Интеллектуальный анализ данных. Weka.

Интеллектуальный анализ данных с помощью Weka

Общее правило экспериментального дизайна - контролировать любые факторы, которые в ваших силах контролировать, и использовать рандомизацию, чтобы обойти проблему факторов, которые вы не можете контролировать.

1. Практикум

В этом *тесте* используется набор данных *contact-lenses.arff* , который был помещен в папку *данных* (в вашей установке Weka) при загрузке Weka. В Weka Explorer откройте набор данных *КОНТАКТНЫХ ЛИНЗ*.

Сколько экземпляров содержится в наборе данных о *контактных линзах*?

Сколько атрибутов содержится в наборе данных о *контактных линзах*?

Сколько возможных значений атрибута *age* ?

Какой из атрибутов имеет значение *уменьшился* ?

1. Практикум

В сфере электроснабжения важно как можно раньше определить будущий спрос на электроэнергию. Если можно будет сделать точные оценки максимальной и минимальной нагрузки для каждого часа, дня, месяца, сезона и года, коммунальные компании смогут значительно сэкономить в таких областях, как установка рабочего резерва, графика технического обслуживания и управление запасами топлива.

- Периодичность электрической нагрузки может проявляться на нескольких основных частотах - очевидна годовая (почему?). А какие другие?
- А как насчет незначительных изменений, которые могут произойти в праздничные дни?
- А как насчет погоды?
- А как насчет общего роста?

2. Практикум

В Weka (Explorer) откройте набор данных *iris.arff*

Это классический набор данных для интеллектуального анализа данных, созданный известным статистиком Р. А. Фишером в 1936 году.

- Какой из атрибутов, взятый сам по себе, хуже всего показывает класс?
- Имеет ли класс *Iris-virginica* склонность к высоким или низким значениям *sepallength*?
- Сколько возможных экземпляров в наборе данных *iris* ?
- Каким значением является атрибут *sepallength* дискретным или числовым?
- Какое минимальное количество атрибутов возможно для создание набора данных и почему?

Лабораторная работа №1

Создание набора данных. Weka.

- Создать набор данных формата ARFF.
- Набор данных должен содержать минимум 3 атрибута.
- У каждого атрибута должно быть минимум два значения при номинальном формате.
- В наборе данных должны быть использованы номинальные и числовые значения.
- В наборе данных должны быть минимум 15 экземпляров.

2. Интеллектуальный анализ данных. Weka.

Набор данных Glass

откройте файл glass.arff

Name:	RI	Type:
Missing:	0 (0%)	Numeric
Distinct:	178	Unique: 145 (68%)
Statistic:	Value	
Minimum:	1.511	
Maximum:	1.534	
Mean:	1.518	
StdDev:	0.008	

Class: Type (Nom) Visualize All

Status: OK

2. Интеллектуальн ый анализ данных. Weka.

Использование классификатора

Используйте J48 для анализа набора данных Glass

- ❖ Откройте файл `glass.arff`
(либо оставьте его открытым с предыдущего урока)
- ❖ Проверьте доступные классификаторы
- ❖ Выберите J48 (`trees>J48`)
- ❖ Запустите
- ❖ Изучите выходные данные
- ❖ Посмотрите на правильно классифицированные экземпляры
... И на the confusion matrix

2. Интеллектуальн ый анализ данных. Weka.

Использование классификатора

Исследуйте J48

- ❖ Откройте панель конфигурации
- ❖ Проверьте дополнительную информацию
- ❖ Изучите варианты
- ❖ Используйте необрезанное дерево
- ❖ Посмотрите на размеры листьев
- ❖ Установите значение **minNumObj** равным 15, чтобы избежать маленьких листьев
- ❖ Визуализируйте дерево с помощью меню

2. Интеллектуальн ый анализ данных. Weka.

Использование классификатора

От C4.5 до J48

- ❖ ID3 (1979)
- ❖ C4.5 (1993)
- ❖ C4.8 (1996?)
- ❖ C5.0 (коммерческий)



J48

2. Интеллектуальн ый анализ данных. Weka.

Использование классификатора

- ❖ классификаторы в Weka
- ❖ классификация набора данных the glass
- ❖ интерпретация выходных данных J48
- ❖ панель конфигурации J48
- ❖ ... вариант: обрезанные или необрезанные деревья
- ❖ ... вариант: избегайте маленьких листьев
- ❖ J48 ~ C4.5

Сбор данных для интеллектуального анализа

Идеальный Датасет – это очищенная выборка без ошибок, выбросов и пропущенных значений, но с полным набором данных, необходимых для решения поставленной задачи.

В реальности мы чаще имеем дело с некорректной, неполной или не достающей информацией.

3 легальных способа сбора чужих данных:

Использование готовых датасетов.

Kaggle - более 50 000 общедоступных наборов данных

Работа с веб-платформами, предоставляющими статистику

Использование информации со сторонних сайтов

Сбора собственных данных:

СБОР ДАННЫХ НА ПРИМЕРЕ СБОРА СУБЪЕКТИВНЫХ ОЦЕНОК.

Базы данных видео со сбором субъективных оценок составляют важную основу для алгоритмов анализа.

Общее правило экспериментального дизайна - контролировать любые факторы, которые в ваших силах контролировать, и использовать рандомизацию, чтобы обойти проблему факторов, которые вы не можете контролировать.

Субъективные тесты

Сборы субъективных оценок на сегодняшний момент.

1. Методология двойной или одинарной непрерывной шкалы качества стимулов
2. Краудсорсинг
3. Пороговые оценки

Субъективные тесты

Основные рекомендации по сбору субъективных оценок:

1. Лабораторная среда
2. Стимулы
3. Участники

3. Практикум

- Откройте набор данных *Glass.arff*. Используйте матрицу неточностей, чтобы определить, сколько экземпляров *headlamps* было ошибочно классифицировано как *build wind float*?
- Откройте набор данных *Labor.arff*, перейдите на панель «Классификация» и запустите классификатор *J48* (с параметрами по умолчанию). Каков процент правильно классифицированных экземпляров?
- Теперь отключите обрезку на панели конфигурации *J48* (набор данных *Labor.arff*), установив для параметра *unpruned* значение *-True*, и запустите его снова. Каков процент правильно классифицированных экземпляров сейчас?
- Постройте вручную дерево решений для созданного набора данных в лабораторной работе №1, проверьте данное решение с помощью *Weka*.

4. Практикум

1. Найти последний документ по Методики субъективной оценки качества телевизионных изображений. Написать название первым пунктом.
2. Определить основные условия лабораторной среды для проведения субъективных тестов. Выписать 2 пунктом.
3. Определить какую информацию должны содержать результаты субъективных тестов при предоставлении в общее пользование. Выписать 3 пунктом.

2. Лабораторная работа

По полученной базе данных определить и выписать 4 пунктом:

- метод сбора информации
- критерии выбора участников
- стимул
- лабораторную среду
- количество последовательностей
- количество последовательностей с артефактами
- недостатки и возможные пути решения

Датасеты для анализа по группам:

LIVE-YT-HFR

LIVE-NFLX-II

LIVE Wild

KoNViD-1k

VideoSet: A large-scale compressed video quality dataset based on JND measurement

Интеллектуальны й анализ данных с помощью Weka. Использование фильтра:

Использование фильтра

Использование фильтра для удаления атрибута

- ❖ Откройте **weather.nominal.arff** (снова!)
- ❖ Проверьте фильтры
 - supervised vs unsupervised
 - attribute vs instance
- ❖ Выберите **unsupervised attribute** фильтр **Remove**
- ❖ Проверьте More information; посмотрите на варианты
- ❖ Установите для **attributeIndices** значение **3** и нажмите OK
- ❖ Примените фильтр
- ❖ Вы можете сохранить результат, нажав Save the result
- ❖ Нажмите Undo

Интеллектуальный анализ данных с помощью Weka.

Использование фильтра

Использование фильтра:

Удалите экземпляры, в которых humidity имеет значение high

- ❖ Supervised или unsupervised?
- ❖ Attribute или instance?
- ❖ Посмотрите на них
- ❖ Выберите **RemoveWithValues**
- ❖ Установите **attributeIndex**
- ❖ Установите **nominalIndices**
- ❖ Нажмите Apply
- ❖ Нажмите Undo

Интеллектуальный анализ данных с помощью Weka. Использование фильтра:

Использование фильтра

Меньше атрибутов, лучше классификация!

- ❖ Откройте файл `glass.arff`
- ❖ Запустите J48 (`trees>J48`)
- ❖ Удалите `Fe`
- ❖ Удалите все атрибуты, кроме `RI` и `MG`
- ❖ Посмотрите на дерево решений

- ❖ Используйте меню для визуализации деревьев решений

Интеллектуальный анализ данных с помощью Weka. Использование фильтра:

Использование фильтра

- ❖ Фильтры в Weka
- ❖ Supervised vs unsupervised, attribute vs instance
- ❖ Чтобы найти подходящий, вам нужно посмотреть!
- ❖ Фильтры могут быть очень мощными
- ❖ Разумное удаление атрибутов может:
 - Повысить производительность
 - Повысить разборчивость

Интеллектуальный анализ данных с помощью Weka. Визуализация данных:

Использование панели Visualize

- Откройте iris.arff
- Вызовите панель «Визуализация»
- Щелкните один из графиков; изучить некоторые примеры
- Нажмите "Цвет класса", чтобы изменить цвет.
- Полоски справа меняются в соответствии с атрибутами: щелкните, чтобы увидеть X ось; щелкните правой кнопкой мыши по ось Y
- Ползунок джиттера
- Показать выбор экземпляра: параметр «Прямоугольник»
- Отправить, сбросить, очистить и сохранить

Интеллектуальный анализ данных с помощью Weka. Визуализация данных:

Будь классификатором!

Интерактивное построение дерева решений

- ❖ Загрузите **segmentchallenge.arff**; посмотрите на набор данных
- ❖ Выберите **UserClassifier** (tree classifier)
- ❖ Используйте набор тестов **segmenttest.arff**
- ❖ Изучите визуализатор данных и визуализатор дерева
- ❖ Постройте **regioncentroidrow** или **intensitymean**
- ❖ Rectangle, Polygon and Polyline selection tools
- ❖ ... several selections ...
- ❖ Нажмите правой кнопкой мыши на **Tree visualizer**, а затем **Accept the tree**

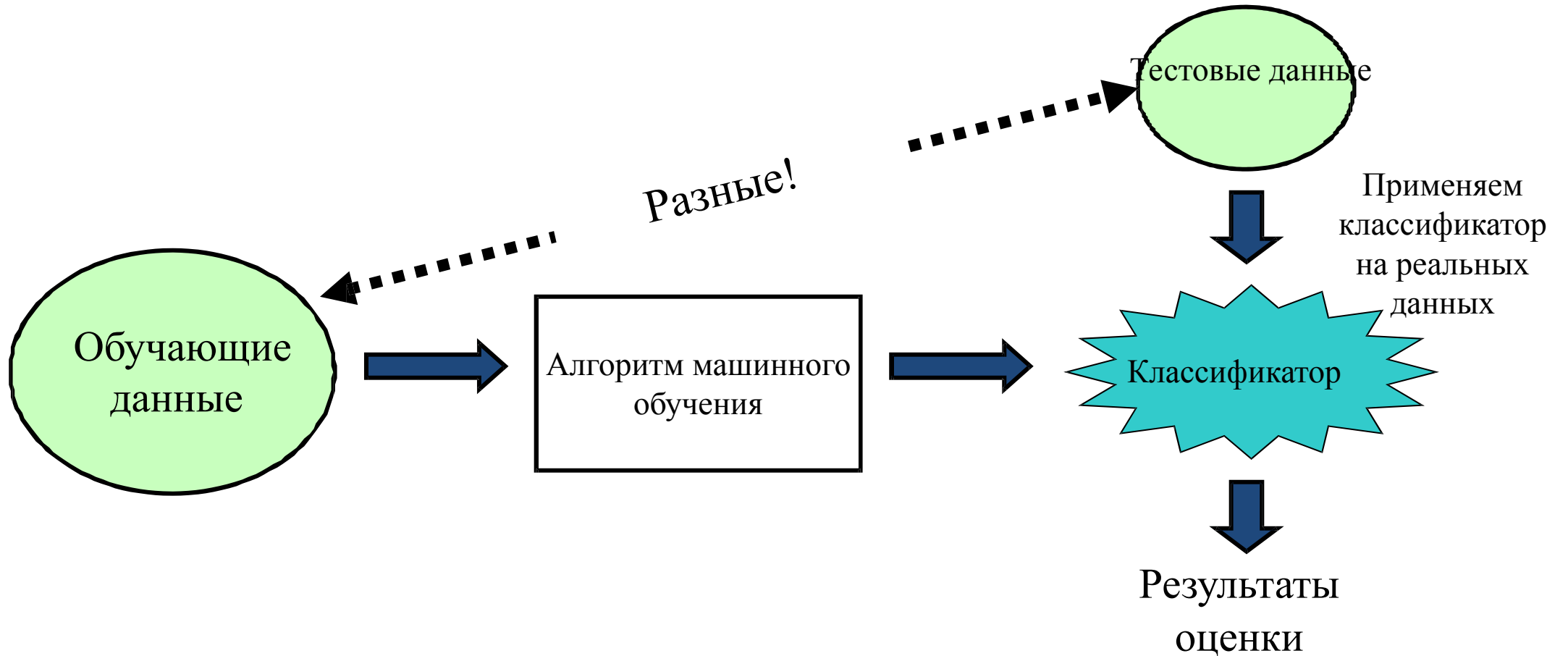
Насколько хорошо вы справляетесь?

Интеллектуальный анализ данных с помощью Weka. Визуализация данных:

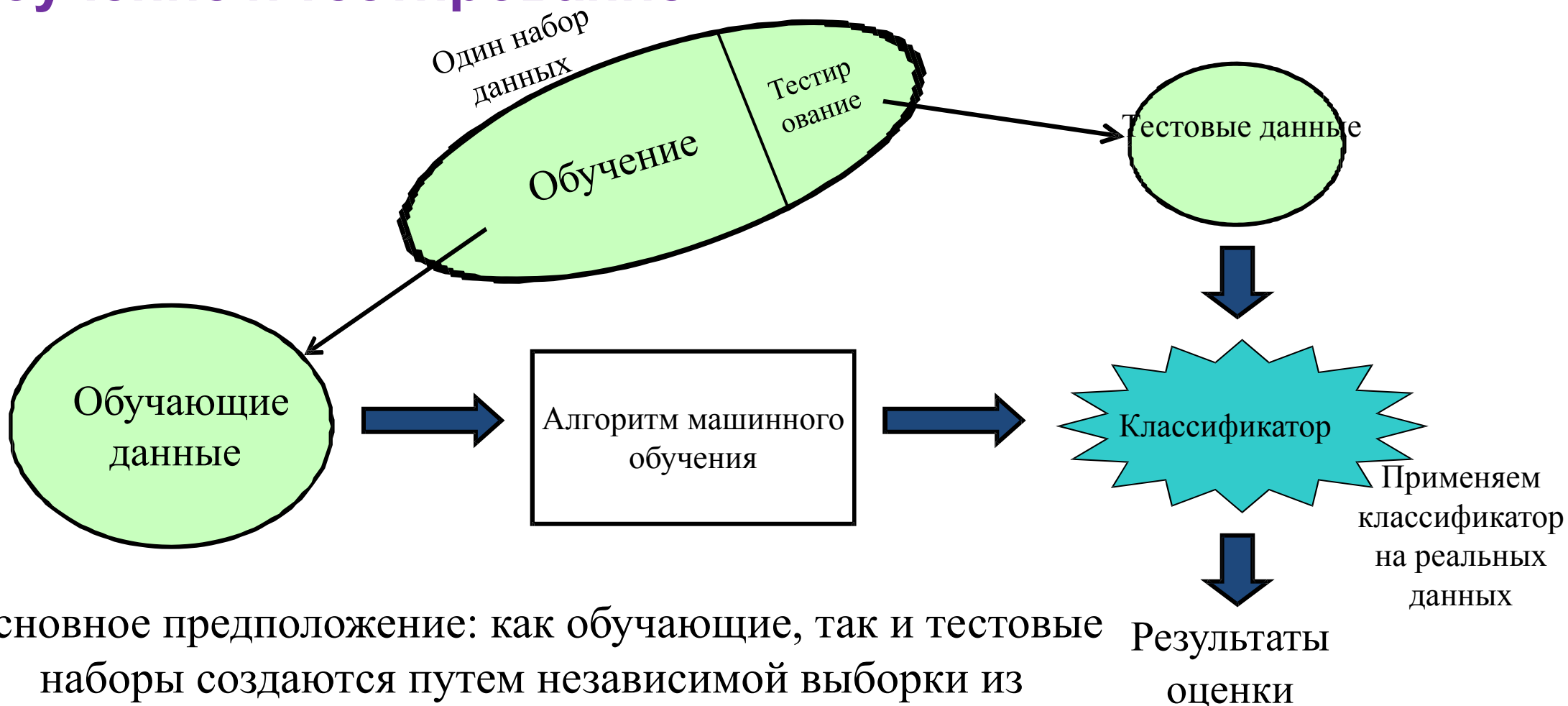
Будь классификатором!

- ❖ Постройте дерево: какую стратегию вы использовали?
- ❖ Если у вас будет достаточно времени, вы сможете создать «идеальное» дерево для набора данных
 - Но будет ли оно хорошо работать с тестовым набором данных?

Обучение и тестирование



Обучение и тестирование



Основное предположение: как обучающие, так и тестовые наборы создаются путем независимой выборки из бесконечной совокупности.

Обучение и тестирование

Используйте J48 для анализа набора данных сегмента

- ❖ Откройте файл **segment-challenge.arff**
- ❖ Выберите дерево решений J48 (**trees>J48**)
- ❖ Выберите прилагаемый тестовый набор **segment-test.arff**
- ❖ Запустите: 96% точности
- ❖ Оцените на тренировочном наборе: 99% точности
- ❖ Оцените по процентному разделению: 95% точности
- ❖ Сделайте это снова: получите точно такой же результат!

Обучение и тестирование

- ❖ Основное предположение:
Как обучающие, так и тестовые наборы создаются путем независимой выборки из бесконечной совокупности
- ❖ Всего один набор данных? — оставьте небольшую часть данных из этого набора для тестирования
- ❖ Мы ожидали бы небольших изменений в результатах
- ❖ ... но Weka каждый раз выдает одни и те же результаты
- ❖ J48 на наборе данных segment-challenge

Повторное обучение и тестирование

Оцените J48 на наборе данных segment-challenge

- ❖ С `segment-challenge.arff` ...
- ❖ и J48 (`trees>J48`)
- ❖ Установите процентное разделение на **90%**
- ❖ Запустите: `9 6.7 %` точности
- ❖ Повторите
- ❖ [**дополнительные параметры**] Повторите с начальными значениями случайного числа 2, 3, 4, 5, 6, 7, 8, 9 10

0.967

0.940

0.940

0.967

0.953

0.967

0.920

0.947

0.933

0.947

Повторное обучение и тестирование

Оцените J48 на наборе данных segment-challenge

Среднее значение выборки	$\bar{x} = \frac{\sum x_i}{n}$	0.967
Дисперсия	$\sigma^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$	0.940
Стандартное отклонение	σ	0.940
		0.967
		0.953
		0.967
		0.920
		0.947
		0.933
		0.947

$$\bar{x} = 0.949, \sigma = 0.018$$

Повторное обучение и тестирование

- ❖ Основное предположение:
обучающие и тестовые наборы, независимо отобранные из
бесконечной совокупности
- ❖ Ожидайте незначительных изменений в результатах...
- ❖ ... получите его, установив начальное значение случайного числа
- ❖ Можно вычислить среднее значение и стандартное отклонение
экспериментально

5. Практическая работа

Откройте набор данных *anneal*

- Сколько атрибутов имеет набор данных *anneal* ?
- Примените неконтролируемый фильтр для атрибутов *-RemoveUseless* . Сколько атрибутов сейчас в наборе данных *anneal* ?
- Определите один из атрибутов, который был удален, нажав кнопку «Отменить», а затем «Применить» . Почему он был убран?

Откройте набор данных *glass.arff* .

- Примените фильтр неконтролируемого атрибута *Нормализовать* . Каков новый диапазон (т.е. минимум и максимум) атрибута *Na* ?

5. Практическая работа

- Отмените действие фильтра Нормализовать и откройте его панель конфигурации. Установите шкалу на 3 и параметр перевода на 1. Снова примените фильтр. Каков диапазон атрибута Na сейчас?
- Отмените изменение и убедитесь, что вы вернулись к исходному набору данных. Теперь примените фильтр неконтролируемых атрибутов «Стандартизировать». Каковы новое среднее значение и стандартное отклонение атрибута K ?
- Снова отмените все изменения в наборе данных стекла. Теперь определите, какой набор атрибутов дает наивысшую точность классификации, используя J48.

6. Практическая работа Поиск неверно классифицированных экземпляров

Откройте набор данных *iris.arff*

- Выберите древовидный классификатор J48 и запустите его (с параметрами по умолчанию). Сколько экземпляров классифицировано неправильно?

- Визуализируйте ошибки классификатора, щелкнув правой кнопкой мыши на список результатов, и используйте визуализацию для определения номеров неправильно классифицированных экземпляров. Какие они?

- Теперь переключите классификатор на *SimpleLogistic*, который вы найдете в категории функций, и запустите его (с параметрами по умолчанию). Сколько экземпляров классифицировано неправильно?

- Какие экземпляры типа *Iris-versicolor* ошибочно классифицируются как *Iris-virginica* ?

7. Практическая работа

Откройте набор данных *segment-challenge.arff*

- Выберите классификатор J48 (параметры по умолчанию), выберите разделение в процентах в качестве параметра теста и определите долю правильно классифицированных экземпляров, когда для размера обучающего набора используются следующие процентные значения: 10%, 20%, 40%. 60%, 80%. Опишите словами закономерность, которую вы наблюдаете?
- Повторите вопрос 1, используя процентное соотношение обучающего набора 90%, 95%, 98% и 99%. Что происходит с количеством правильно классифицированных экземпляров и почему?
- Повторение вопроса 1 с процентным соотношением обучающей выборки 99% дает цифру 100% точности на тестовой выборке. Означает ли это, что это создает идеальный классификатор для проблемы сегментации и почему?



7. Практическая работа

- Основываясь на вышеупомянутых экспериментах, какова ваша наилучшая оценка истинной точности J48 в наборе данных *проблем сегмента* ?
- Какая вероятность того, что J48 не сделает ошибок на 15 независимо выбранных тестовых экземплярах, если его точность для каждого экземпляра составляет 95% и почему (с доказательством, используя математику)?
- Верно ли утверждение, что «чем больше тестовых данных, тем выше вероятность успеха классификатора» ? Объяснить ответ.
- Когда для оценки используется опция *процентного разделения* , насколько хороша производительность, если (а) почти никакие данные не используются для тестирования; (б) почти все данные используются для тестирования? И почему?

8. Практическая работа

Откройте набор данных *diabetes.arff*

- Выберите *процентное разделение* в качестве параметра теста и установите *процентное соотношение для обучения 80%*. Сколько экземпляров будет использовано для обучения, а сколько - для тестирования? И почему?
- Выберите классификатор J48 (параметры по умолчанию) и оцените его со следующими начальными значениями (*дополнительные параметры*): 1, 2, 3, 4, 5. Укажите минимальные и максимальные значения количества неправильно классифицированных экземпляров?
- Какое среднее значение точности для этих пяти начальных значений? Объяснить ответ.
- Какое стандартное отклонение точности для этих пяти значений? И почему? Объяснить ответ, используя математику.
- Если бы вы провели эксперимент с 10 различными случайными начальными числами, а не с 5, как вы ожидаете, это повлияет на среднее значение и стандартное отклонение? Объяснить ответ.

Лабораторная работа 3

Откройте свой набор данных.

- *Выберите древовидный классификатор J48 и запустите его (с параметрами по умолчанию). Сколько экземпляров классифицировано неправильно?*
- *Визуализируйте ошибки классификатора, щелкнув правой кнопкой мыши список результатов, и используйте визуализацию для определения номеров экземпляров неправильно классифицированных экземпляров. Какие они?*

А как насчет объяснения (вашему партнеру, братьям и сестрам, родителям или детям)... *каково это - заниматься интеллектуальным анализом данных?*

Лабораторная работа 3

Какая максимальная точность, которую можно достичь с помощью *UserClassifier* ? Указать число и объяснить почему.

Объясните почему изменение начального числа случайных чисел в Weka Explorer приводит к получению другого результата?

Объясните почему Weka использует генератор случайных чисел (простую небольшую программу), но каждый раз генерирует одну и ту же последовательность?

Базовая точность

Используйте набор данных о диабете и задержку по умолчанию

- ❖ Откройте файл **diabetes.arff**
- ❖ Выберите вариант тестирования: Процентное разделение
- ❖ Попробуйте следующие классификаторы:
 - **trees > J48** 76
 - **bayes > NaiveBayes** %
 - **lazy > IBk** 77
 - **rules > PART** %
- (мы изучим их позже) 73
- ❖ 768 экземпляров (500 отрицательных, 268 положительных) %
- ❖ Всегда угадывает наиболее популярный класс “отрицательный”: 74
- 500/768 65% %
- ❖ **rules > ZeroR**: наиболее вероятный класс!

Базовая точность

Иногда простые методы лучше!

- ❖ Откройте файл `supermarket.arff` и слепо примените

rules > *ZeroR* 64%

trees > *J48* 63%

bayes > *NaiveBayes* 63%

lazy > *IBk* 38%

(!!)

- ❖ *rules* > *PART* 63%

Атрибуты не являются информативными

- ❖ Не просто применяйте Weka к набору данных:
нужно понимать, что происходит!!

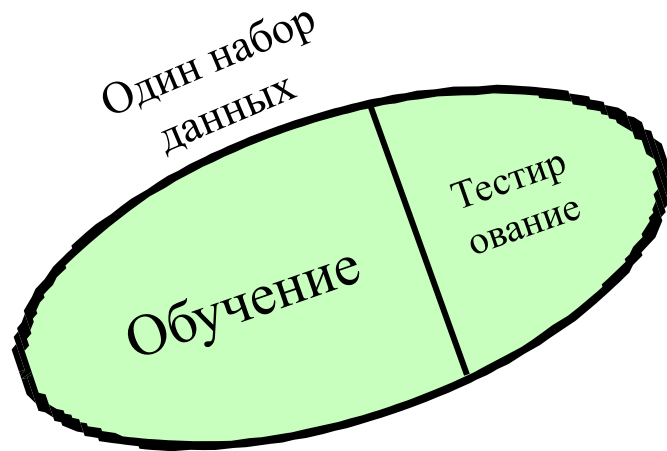
Базовая точность

- ❖ Подумайте, могут ли различия быть значительными
- ❖ Всегда старайтесь придерживаться простой базы, например **rules > ZeroR**
- ❖ Посмотрите на набор данных
- ❖ Не применяйте Weka слепо: попытайтесь понять, что происходит!

Базовая точность

- ❖ Можем ли мы улучшить ситуацию с повторной задержкой? (т.е. уменьшить дисперсию)
- ❖ Перекрестная проверка
- ❖ Стратифицированная перекрестная проверка

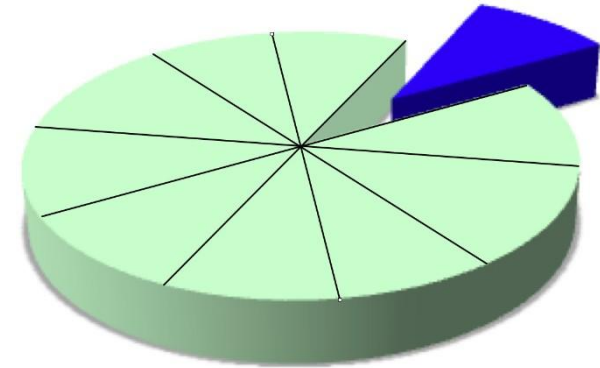
- ❖ Повторная задержка
(оставляем 10% для тестирования, повторяем 10 раз)



Перекрестная проверка

10-кратная перекрестная проверка

- ❖ Разделите набор данных на 10 частей
- ❖ Каждую часть по очереди оставляйте для тестирования
- ❖ Усредните результаты
- ❖ Каждая часть данных использовалась один раз для тестирования, 9 раз для обучения

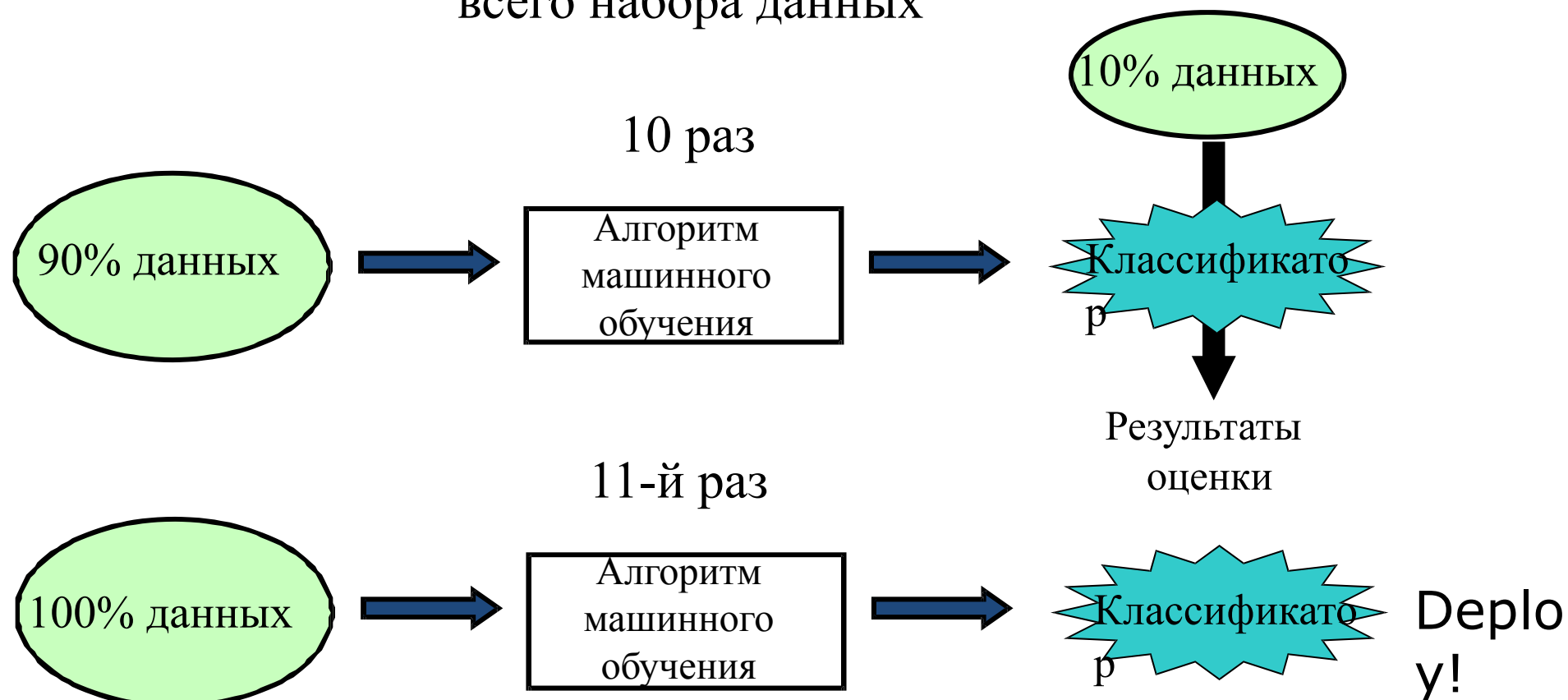


Стратифицированная перекрестная проверка

- ❖ Убедитесь, что каждая часть имеет правильную пропорцию значения каждого класса

Перекрестная проверка

После перекрестной проверки Weka выводит дополнительную модель, построенную на основе всего набора данных



Перекрестная проверка

- ❖ Перекрестная проверка лучше, чем повторная задержка
- ❖ Стратифицированная еще лучше
- ❖ При 10-кратной перекрестной проверке Weka 11 раз вызывает алгоритм обучения
- ❖ Практическое эмпирическое правило:
 - Много данных? – используйте процентное разделение
 - В других случаях стратифицированную 10—кратную перекрестную проверку

Результаты перекрестной проверки

Действительно ли перекрестная проверка лучше, чем повторная задержка?

- ❖ Набор данных **diabetes**
- ❖ Базовая точность (**rules > ZeroR**): 65.1%
- ❖ **trees > J48**
- ❖ 10-кратная перекрестная проверка 73.8%
- ❖ ... с разными начальными значениями случайных чисел

	1	2	3	4	5	6	7	8	9	10
73.8	75.0	75.5	75.5	74.4	75.6	73.6	74.0			
			74.5	73.0						

Результаты перекрестной проверки

	Задержка (10%)	Перекрестная проверка (10-кратная)
Среднее значение выборки $\bar{x} = \frac{\sum x_i}{n}$	75.3	73.8
Дисперсия $\sigma^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$	77.9	75.0
Стандартное отклонение σ deviation	80.5	75.5
	74.0	75.5
	71.4	74.4
	70.1	75.6
	79.2	73.6
	71.4	74.0
	80.5	74.5
	67.5	73.0
	$\bar{X} =$	$\bar{X} =$
	74.8	74.5
	$\sigma =$	$\sigma =$
	4.6	0.9

Результаты перекрестной проверки

- ❖ Почему 10-кратная? Если 20-кратная: 75.1%
- ❖ Перекрестная проверка действительно лучше, чем повторная задержка
- ❖ Это уменьшает дисперсию оценки

Простота прежде всего!

Простые алгоритмы часто работают очень хорошо!

- ❖ Существует много видов простой структуры, например:
 - *Один атрибут выполняет всю работу*
 - *Атрибуты вносят равный и независимый вклад*
 - *Дерево решений, которое проверяет несколько атрибутов*
 - *Вычислить расстояние от обучающих экземпляров*
 - *Результат зависит от линейной комбинации атрибутов*

- ❖ Успех метода зависит от предметной области
 - *Интеллектуальный анализ данных - это экспериментальная наука*

Простота прежде всего!

OneR: Один атрибут выполняет всю работу

- ❖ 1-уровневое “дерево решений”
 - *т.е. правила, которые проверяют один конкретный атрибут*

- ❖ Основной вариант
 - *Одна ветвь для каждого значения*
 - *Каждой ветви присваивается наиболее частый класс*
 - *Частота ошибок: доля экземпляров, которые не принадлежат к классу большинства соответствующей ветви*
 - *Выбирается атрибут с наименьшей частотой ошибок*

Простота прежде всего!

Для каждого значения атрибута, создайте правило следующим образом:

подсчитайте, как часто появляется каждый класс

найдите наиболее частый класс

создайте правило, присваивающее этому классу значение

атрибута

Рассчитайте частоту ошибок правил этого атрибута. Выберите атрибут с наименьшей частотой ошибок.

Простота прежде всего!

Outlook	Temp	Humidity	Wind	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Sunny	Mild	High	False	No
Sunny	Cool	Normal	False	Yes
Rainy	Mild	Normal	False	Yes
Sunny	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Rainy	Mild	High	True	No

Attribute	Rules	Errors	Total errors
Outlook	Sunny → No	2/5	4/14
	Overcast → Yes	0/4	
	Rainy → Yes	2/5	
Temp	Hot → No*	2/4	5/14
	Mild → Yes	2/6	
	Cool → Yes	1/4	
Humidity	High → No	3/7	4/14
	Normal → Yes	1/7	
Wind	False → Yes	2/8	5/14
	True → No*	3/6	

* указывает на ничью

Простота прежде всего!

Используйте OneR

- ❖ Откройте [weather.nominal.arff](#)
- ❖ Выберите OneR (**rules>OneR**)
- ❖ Посмотрите на правило (*примечание: Weka выполняет OneR 11 раз*)

Простота прежде всего!

OneR: Один атрибут выполняет всю работу

❖ Невероятно простой метод, описанный в 1993 году

“Очень простые правила классификации хорошо работают с наиболее часто используемыми наборами данных”

– Экспериментальная оценка на 16 наборах данных

– Используется перекрестная проверка

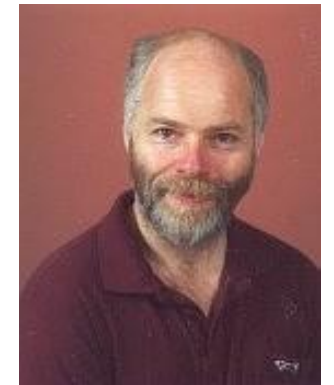
– Простые правила часто превосходили гораздо более сложные методы

❖ Как это может так хорошо работать?

– некоторые наборы данных действительно просты

– некоторые из них настолько малы / шумны / сложны, что у них ничему нельзя научиться!

Rob Holte,
Alberta, Canada



9. Практическая работа

Проверьте, что случайное начальное число значения по умолчанию 1, прежде чем приступить к практикам.

Iris.arff набор данных состоит из трех классов (*Iris-setosa*, *Iris-лишай*, *Iris-virginica*), с 50 экземпляров каждого.

- Какая точность ZeroR для этого набора данных при тестировании на обучающем наборе и какая степень успеха?
- Как в данном случае работает ZeroR?
- На практике, какой процент успеха ZeroR для набора данных радужной оболочки глаза при оценке с использованием процентного разделения по умолчанию (66%) ?
- Почему могут существовать некоторое статистическое отклонение от ожидаемого значения?

9. Практическая работа

Проверьте, что случайное начальное число значения по умолчанию 1, прежде чем приступить к практикам.

Откройте набор данных *segment-challenge.arff*, перейдите на вкладку Classify. Выберите классификатор J48 (параметры по умолчанию), выберите перекрестную проверку в качестве параметра теста, используя 10 крат. Оцените J48 со следующими случайными начальными значениями: 11, 12, 13, 14, 15.

- Какое среднее значение точности со случайными начальными числами 11, 12, 13, 14 и 15?
- Какое стандартное отклонение точности?
- Когда вы провели описанный выше эксперимент, сколько раз Weka запускала алгоритм J48?

Для того же набора данных выберите Процентное разделение в качестве параметра теста с 90% в качестве параметра.

Оцените J48 с теми же начальными значениями, что и раньше: 11, 12, 13, 14, 15

- Какая средняя точность?
- Какое стандартное отклонение точности?
- Когда вы проводили описанный выше эксперимент, сколько раз Weka выполняла алгоритм J48 для создания дерева решений и почему?



10. Практическая работа

Откройте набор данных *iris.arff* и перейдите на вкладку Classify . Выполните 10-кратную перекрестную проверку с помощью ZeroR и OneR.

- Какой классификатор обеспечивает более высокую точность?
- Какой атрибут использует OneR для создания правила в предыдущем эксперименте при использовании полного набора данных?
- Может ли быть набор данных, по которому ZeroR превосходит OneR и почему?
- Может ли быть набор данных, для которого ZeroR превосходит OneR при оценке на данных обучения? Почему, предоставьте проверку используя математическую индукцию (подсказка пример 2-х классового случая с классами «да» и «нет»)?

Лабораторная работа 4

Откройте набор данных *iris.arff*

- Оцените точность базового метода ZeroR, используя перекрестную проверку с 10, 11, 12, 13, 14 и 15 кратностями.
- Какие минимальное и максимальное значение результатов, полученных с помощью ZeroR для набора данных радужной оболочки глаза с использованием перекрестной проверки с 10, 11, 12, 13, 14 и 15 кратностями?
- Все значения, полученные в предыдущем вопросе, были меньше или равны истинному значению точности ZeroR в 33% в этом наборе данных. Это совпадение? Почему?

Лабораторная работа 4

Предположим, что точность ZeroR для набора данных *iris.arff* оценивалась с использованием перекрестной проверки с 5, 10 и 25 кратностями.

Какую точность вы ожидаете, не проводя эксперимента и почему (объяснить, используя цифры)?

Какая вероятность успеха ZeroR на наборе данных *iris.arff*, если оценивать его с помощью 150-кратной перекрестной проверки? Сначала хорошенько подумайте об этом и объясните, а затем подтвердите свой ответ с помощью Weka.

Как вы оцениваете работу классификатора? Попробуйте объяснить (своему партнеру, братьям и сестрам, родителям или детям), как оценивать эффективность системы обучения, если вы даже не знаете, на каких данных она будет использоваться. Сможете ли вы убедить их, почему оценивать его на данных, используемых для обучения, - это абсолютно ужасная идея?

Переобучение

- ❖ Любой метод машинного обучения может “переобучать” обучающие данные ...
 - ... путем создания классификатора, который слишком точно соответствует данным обучения
- ❖ Хорошо работает с обучающими данными, но не с данными независимых тестов
- ❖ Помните “Пользовательский классификатор”? Представьте себе утомительное нанесение крошечного круга вокруг каждой отдельной точки данных обучения
- ❖ Переобучение - это общая проблема
- ❖ ... мы продемонстрируем это с помощью OneR

Переобучение

Числовые атрибуты

Outlook	Temp	Humidity	Wind	Play
Sunny	85	85	False	No
Sunny	80	90	True	No
Overcast	83	86	False	Yes
Rainy	75	80	False	Yes
...

Attribute	Rules	Errors	Total errors
Temp	85 → No	0/1	0/14
	80 → Yes	0/1	
	83 → Yes	0/1	
	75 → No	0/1	
...		...	

- ❖ У OneR есть параметр, который ограничивает сложность таких правил

Переобучение

Поэкспериментируйте с OneR

- ❖ Откройте файл `weather.numeric.arff`
- ❖ Выберите OneR (`rules>OneR`)
- ❖ Результирующее правило основано на атрибуте `outlook`, так что удалите `outlook`
- ❖ Правило основано на атрибуте `humidity`

```
humidity: < 82.5 -> yes  
         >= 82.5 -> no
```

(10/14 правильных экземпляров)

Переобучение

Поэкспериментируйте с набором данных diabetes

- ❖ Откройте файл **diabetes.arff**
- ❖ Выберите ZeroR (**rules>ZeroR**)
- ❖ Используйте перекрестную проверку: 65.1%
- ❖ Выберите OneR (**rules>OneR**)
- ❖ Используйте перекрестную проверку: 72.1%
- ❖ Посмотрите на правило (plas = plasma glucose concentration, концентрация глюкозы в плазме крови)
- ❖ Измените параметр **minBucketSize** на **1** : 54.9%
- ❖ Оцените на тренировочном наборе : 86.6%
- ❖ Посмотрите на правило еще раз

Переобучение

- ❖ Переобучение — это общее явление, от которого страдают все методы машинного обучения
- ❖ Это одна из причин, почему вы никогда не должны оценивать на тренировочном наборе
- ❖ Переобучение может происходить в более общем случае
- ❖ Например, попробуйте множество методов машинного обучения, выберите лучший для ваших данных
 - — вы не можете ожидать такой же производительности на новых тестовых данных
- ❖ Правило: Разделять данные на обучающие, тестовые, проверочные наборы.

Использование вероятностей

(OneR: Один атрибут выполняет всю работу)

Противоположная стратегия: используйте все атрибуты **“Наивный Байесовский” метод**

- ❖ Два предположения: Атрибуты
 - *одинаково важно априори*
 - *статически независимы (с учетом значения класса)*
т.е., знание значения одного атрибута ничего не говорит о значении другого
(если известен класс)
- ❖ Предположение о независимости никогда не бывает правильным!
- ❖ Но ... часто хорошо работает на практике

Использование вероятностей

Вероятность события H при наличии свидетельства E

$$\Pr[H \mid E] = \frac{\Pr[E \mid H] \Pr[H]}{\Pr[E]}$$

class
instance

- ❖ $\Pr[H]$ априорная вероятность H
 - Вероятность события до появления доказательств
- ❖ $\Pr[H \mid E]$ апостериорная вероятность H
 - Вероятность события после того, как увидят доказательства
- ❖ “Наивное” предположение:
 - Доказательства распадаются на части, которые являются независимыми



$$\Pr[H \mid E] = \frac{\Pr[E_1 \mid H] \Pr[E_2 \mid H] \dots \Pr[E_n \mid H] \Pr[H]}{\Pr[E]}$$

Томас Байес, британский математик, 1702 –1761

Использование вероятностей

Outlook	Temperature		Humidity		Wind		Play						
	Yes	No	Yes	No	Yes	No	Yes	No					
Sunny	2	3	Hot	2	2	High	3	4	False	6	2	9	5
Overcast	4	0	Mild	4	2	Normal	6	1	True	3	3		
Rainy	3	2	Cool	3	1								
Sunny	2/9	3/5	Hot	2/9	2/5	High	3/9	4/5	False	6/9	2/5	9/14	5/14
Overcast	4/9	0/5	Mild	4/9	2/5	Normal	6/9	1/5	True	3/9	3/5		
Rainy	3/9	2/5	Cool	3/9	1/5								

Outlook	Temp	Humidity	Wind	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Sunny	Mild	High	False	No
Sunny	Cool	Normal	False	Yes
Rainy	Mild	Normal	False	Yes
Sunny	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Rainy	Mild	High	True	No

$$\Pr[H | E] = \frac{\Pr[E_1 | H] \Pr[E_2 | H] \dots \Pr[E_n | H] \Pr[H]}{\Pr[E]}$$

Использование вероятностей

	Outlook		Temperature		Humidity		Wind		Play				
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No			
Sunny	2	3	Hot	2	2	High	3	4	False	6	2	9	5
Overcast	4	0	Mild	4	2	Normal	6	1	True	3	3		
Rainy	3	2	Cool	3	1								
Sunny	2/9	3/5	Hot	2/9	2/5	High	3/9	4/5	False	6/9	2/5	9/14	5/14
Overcast	4/9	0/5	Mild	4/9	2/5	Normal	6/9	1/5	True	3/9	3/5		
Rainy	3/9	2/5	Cool	3/9	1/5								

Новый день:

Outlook	Temp.	Humidity	Wind	Play
Sunny	Cool	High	True	?

$$\Pr[H | E] = \frac{\Pr[E_1 | H] \Pr[E_2 | H] \dots \Pr[E_n | H] \Pr[H]}{\Pr[E]}$$

Likelihood of the two classes

For "yes" = $2/9 \times 3/9 \times 3/9 \times 3/9 \times 9/14 = 0.0053$

For "no" = $3/5 \times 1/5 \times 4/5 \times 3/5 \times 5/14 = 0.0206$

Conversion into a probability by normalization:

$P(\text{"yes"}) = 0.0053 / (0.0053 + 0.0206) = 0.205$

$P(\text{"no"}) = 0.0206 / (0.0053 + 0.0206) = 0.795$

Использование вероятностей

Outlook	Temp.	Humidity	Wind	Play
Sunny	Cool	High	True	?

Доказательство E

Вероятность
класса "yes"

$$\begin{aligned}
 \Pr[\text{yes} | E] &= \Pr[\text{Outlook} = \text{Sunny} | \text{yes}] \\
 &\quad \times \Pr[\text{Temperature} = \text{Cool} | \text{yes}] \\
 &\quad \times \Pr[\text{Humidity} = \text{High} | \text{yes}] \\
 &\quad \times \Pr[\text{Windy} = \text{True} | \text{yes}] \\
 &\quad \times \frac{\Pr[\text{yes}]}{\Pr[E]} \\
 &= \frac{\frac{2}{9} \times \frac{3}{9} \times \frac{3}{9} \times \frac{3}{9} \times \frac{9}{14}}{\Pr[E]}
 \end{aligned}$$

Использование вероятностей

Используйте Наивный Байесовский метод

- ❖ Откройте файл `weather.nominal.arff`
- ❖ Выберите Наивный Байесовский метод (`bayes>NaiveBayes`)
- ❖ Посмотрите на классификатор
- ❖ Избегайте нулевых частот: начните все отсчеты с 1

Использование вероятностей

- ❖ “Naïve Bayes”: все атрибуты вносят одинаковый и независимый вклад
- ❖ На удивление хорошо работает
 - даже если предположение о независимости явно нарушено
- ❖ Почему?
 - классификация не требует точных оценок вероятности
до тех пор, пока наибольшая вероятность относится к правильному классу
- ❖ Добавление избыточных атрибутов вызывает проблемы (например, одинаковые атрибуты) → *выбор атрибута*



11. Практическая работа

Откройте `weather.numeric.arff` набор данных и проверьте данные с помощью Edit кнопки Weka в Preprocess панели.

- Какая максимальная точность правил, основанных на температуре и влажности соответственно, с точки зрения количества правильно спрогнозированных обучающих примеров?
- В следующих вопросах исследуется влияние параметра OneR `minBucketSize` на производительность и сложность правил путем создания графиков, где `minBucketSize` находится в диапазоне от 1 до 10.

Откройте набор данных `glass.arff`, перейдите на вкладку «Классификация» и выберите OneR. Нарисуйте график точности данных обучения (по вертикальной оси) по сравнению с `minBucketSize` (по горизонтальной оси). Опишите.

Создайте график перекрестной проверки точности по `minBucketSize`. Опишите.



11. Практическая работа

Рассмотрите сложность правила, которое генерирует OneR, измеряемое его размером - количеством тестов, которые оно включает.

Будет ли сложность правила в Weka зависеть от того, используется ли обучающий набор или перекрестная проверка для оценки? Объясните.

Начертите размер созданного правила относительно minBucketSize . Меню «More options» на панели «Классификация» можно использовать для настройки вывода. В зависимости от настройки Weka сгенерирует один или несколько разделов.



12. Практическая работа

Откройте набор данных *vote.arff* и выберите классификатор NaiveBayes с параметрами по умолчанию и 10-кратной перекрестной проверкой в качестве метода оценки. Это исторический набор данных, взятый из базы данных записей голосования Конгресса США за 1984 год.

- Какая точность NaiveBayes в этом наборе данных?
- Вернитесь на вкладку « Предварительная обработка » и скопируйте 12-й атрибут, «расходы на образование» , десять раз, используя фильтр «Копировать». Какая точность NaiveBayes в этом новом наборе данных, снова оцененном с помощью 10-кратной перекрестной проверки?
- Вернитесь на вкладку Preprocess и скопируйте тот же атрибут еще десять раз. Какая точность сейчас?



12.

Практическая работа

Вы, вероятно, думаете, что если бы вы продолжали копировать атрибут «расходы на образование» и оценивали его с помощью 10 -кратной перекрестной проверки, точность постепенно снижалась бы, пока, наконец, не выровнялась. И это правильно!

При какой процентной точности это выравнивается? Объясните, используя байсовский подход.

Если точность наивного Байеса постоянно ухудшается по мере добавления копий определенного атрибута (как это происходит здесь для расходов на образование), как вы думаете, улучшится ли это в данном случае, если этот атрибут будет полностью удален из набора данных?

Лабораторная работа 9

Откройте набор данных *breast-cancer.arff* в текстовом редакторе и прочтите комментарии в начале, чтобы ознакомиться с данными, типами атрибутов и другой информацией об атрибутах.

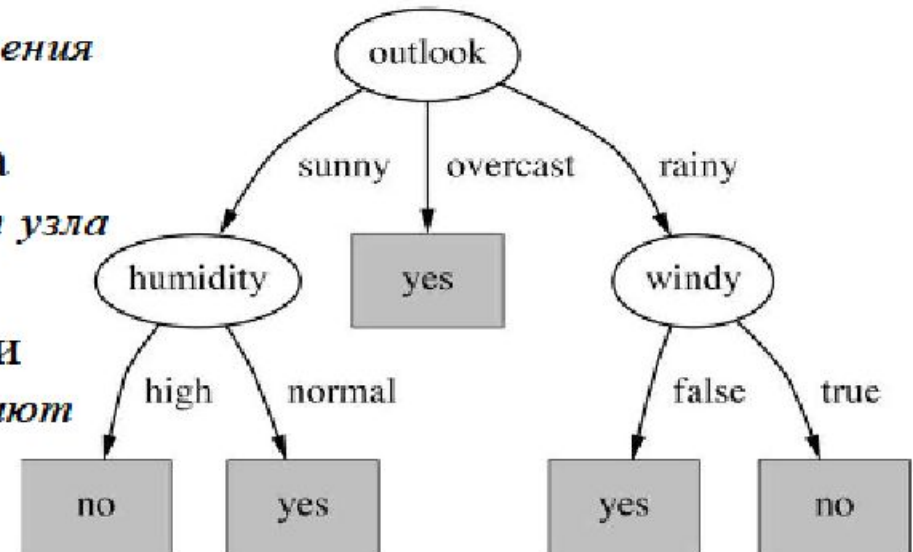
- Набор данных был создан Институтом онкологии в Любляне. Для какого еще исследования они внесли свой вклад?
- Просматривая комментарии в файле ARFF, определите, сколько возможных значений существует для атрибута возраста и сколько из этих значений используется в наборе данных.

Мы приглашаем вас обсудить идею вероятности, гипотезу, основанную на доказательствах, априорную и апостериорную вероятность и что на самом деле означает «наивное» предположение.

Деревья решений

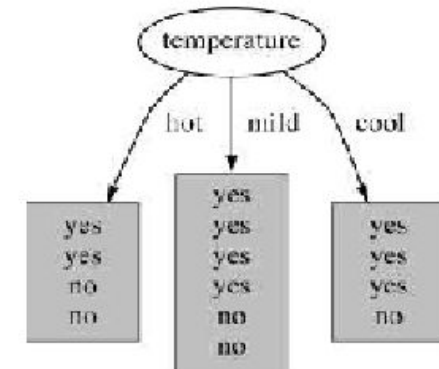
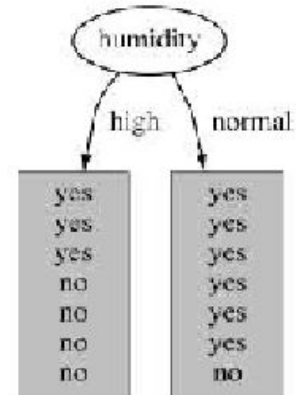
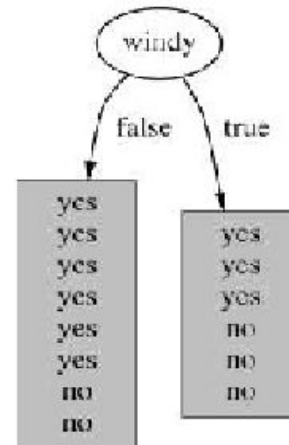
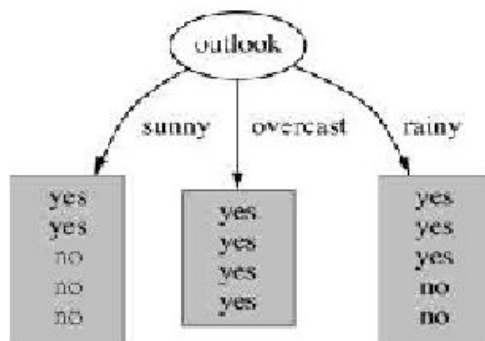
Сверху-вниз: рекурсивный принцип «разделяй и властвуй»

- ❖ **Выбрать атрибут для корневого узла**
 - Создать ветку для каждого возможного значения атрибута
- ❖ **Разделить экземпляры на подмножества**
 - По одному на каждую ветвь, отходящую от узла
- ❖ **Повторить рекурсивно для каждой ветки**
 - используя только экземпляры, которые достигают ветки
- ❖ **Остановитесь**
 - если все экземпляры имеют один и тот же класс



Деревья решений

Какой атрибут выбрать?



Деревья решений

Какой атрибут лучший?

- ❖ Цель: получить самое маленькое дерево
- ❖ Эвристика
 - выберите атрибут, который создает «самые чистые узлы»
 - т.е. наибольший информационный прирост
- ❖ Теория информации: измерение информации в битах

$$\text{энтропия } (p_1, p_2, \dots, p_n) = -p_1 \log p_1 - p_2 \log p_2 \dots - p_n \log p_n$$

Прирост информации

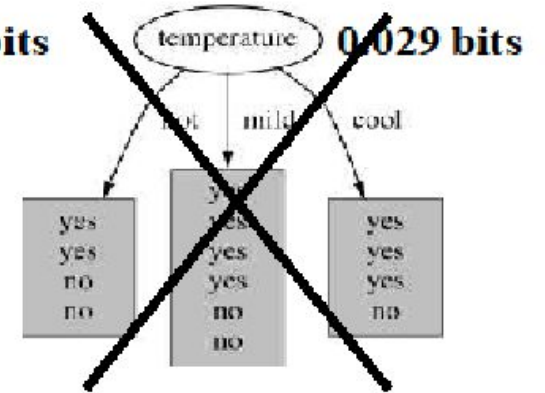
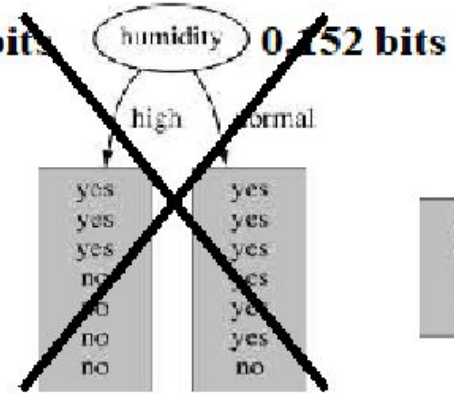
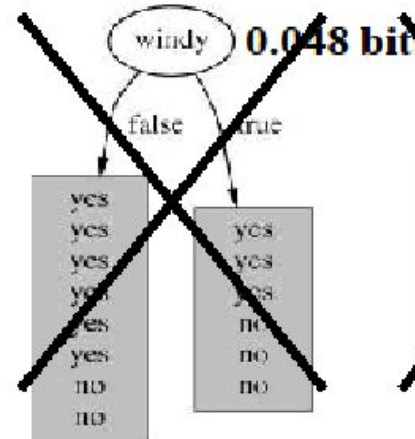
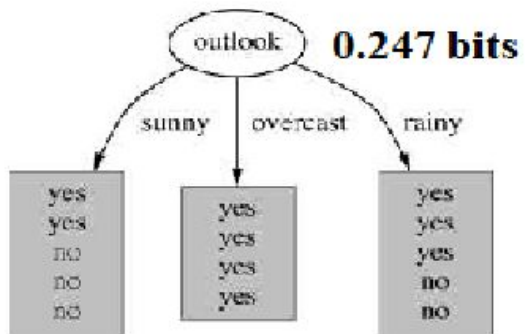
- Количество информации, полученной при знании значения атрибута
- (Энтропия распределения до разделения) – (энтропия распределения после него)



Клод Шеннон, американский математик и ученый 1916–2001

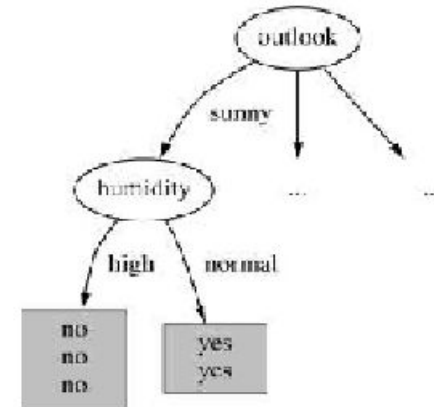
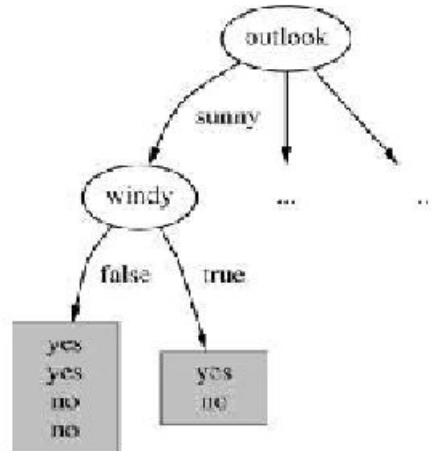
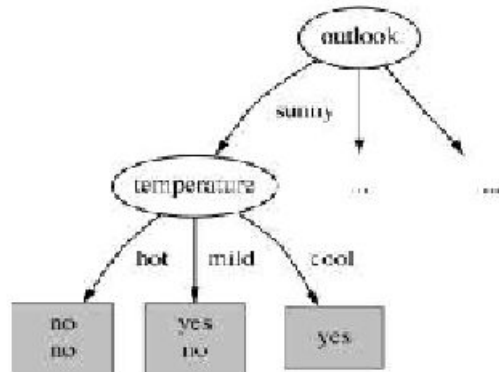
Деревья решений

Какой атрибут выбрать?



Деревья решений

Продолжайте разделять ...



$$\text{gain}(\text{temperature}) = 0.571 \text{ bits}$$

$$\text{gain}(\text{windy}) = 0.020 \text{ bits}$$

$$\text{gain}(\text{humidity}) = 0.971 \text{ bits}$$

Деревья решений

Используйте J48 для данных о погоде

- ❖ Откройте файл **weather.nominal.arff**
- ❖ Выберите J48 устройство обучения дерева решений (**деревья>J48**)
- ❖ Посмотрите на дерево
- ❖ Используйте контекстное меню, чтобы визуализировать дерево

Деревья решений

- ❖ **J48: “индукция деревьев решений сверху-вниз”**
- ❖ Надежно основан на теории информации
- ❖ Производит дерево, которое люди могут понять
- ❖ Множество различных критериев для выбора атрибута
 - редко имеют большое значение
- ❖ Нуждается в дальнейшей модификации, чтобы быть полезным на практике
 - (следующий урок)

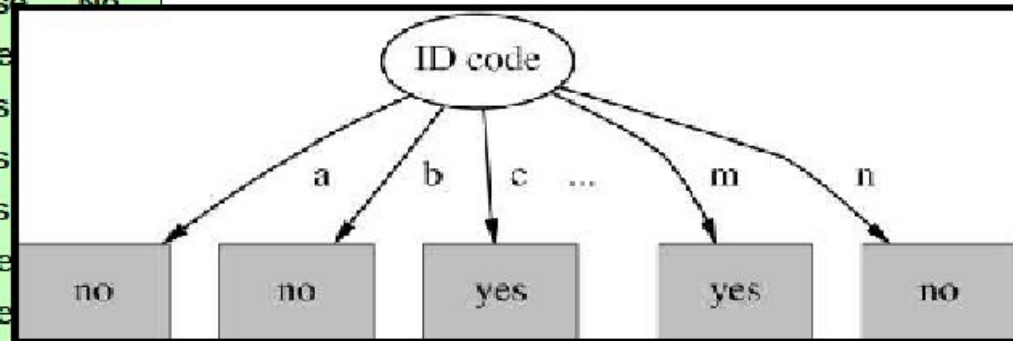
Упрощение деревьев решений



Упрощение деревьев решений

Сильно разветвленные атрибуты — Крайний случай: ID код

ID code	Outlook	Temp	Humidity	Wind	Play
a	Sunny	Hot	High	False	No
b	Sunny	Hot	High	True	No
c	Overcast	Hot	High	False	No
d	Rainy	Mild	High	False	No
e	Rainy	Cool	Normal	False	No
f	Rainy	Cool	Normal	True	No
g	Overcast	Cool	Normal	True	No
h	Sunny	Mild	High	False	No
i	Sunny	Cool	Normal	False	Yes
j	Rainy	Mild	Normal	False	Yes
k	Sunny	Mild	Normal	True	Yes
l	Overcast	Mild	High	True	Yes
m	Overcast	Hot	Normal	False	Yes
n	Rainy	Mild	High	True	No



Прирост информации
максимальный
(0.940 bits)

Упрощение деревьев решений

Как упростить?

- ❖ Не продолжайте разбиение, если узлы становятся очень маленькими (параметр J48 `minNumObj`, значение по умолчанию 2)
- ❖ Построить полное дерево, а затем продолжить работу с листьями, применяя статический тест на каждом этапе (параметр `confidenceFactor`, значение по умолчанию 0.25)
- ❖ Иногда полезно упрощать внутренний узел, подняв поддереву под ним на один уровень выше (`subtreeRaising`, по умолчанию `true`)
- ❖ Беспорядочно ... сложно ... не особо понятно

Упрощение деревьев решений

Переоснащение (снова!)

Иногда упрощение дерева решений дает лучшие результаты

- ❖ Откройте файл **diabetes.arff**
- ❖ Выберите J48 устройство обучения дерева решений (**trees>J48**)
- ❖ Упрощения по умолчанию: 73.8% accuracy, tree has 20 leaves, 39 nodes
- ❖ Отключить упрощение: 72.7% accuracy, 22 leaves, 43 nodes
- ❖ Экстремальный пример: **breast-cancer.arff**
- ❖ По умолчанию (упрощено): 75.5% accuracy, tree has 4 leaves, 6 nodes
- ❖ Неупрощенный: 69.6% accuracy, 152 leaves, 179 nodes

Упрощение деревьев решений

- ❖ C4.5/J48 – популярный ранний метод машинного обучения
- ❖ Много разных способов упрощения
 - в основном изменить размер упрощенного дерева
- ❖ Упрощение – это общий метод, который может применяться к структурам, отличным от деревьев (например, к правилам принятия решений).
- ❖ Одномерные vs. Многомерные деревья решений
 - Одиночные vs. составные тесты на узлах
- ❖ От C4.5 до J48



Росс Квинлан, Австралийский ученый-компьютерщик



13. Практическая работа

Это задание посвящено деревьям решений и алгоритму J48. Мы уже использовали J48 много раз, поэтому вместо того, чтобы делать больше, давайте воспользуемся этой возможностью, чтобы поближе взглянуть на выходные данные, которые Weka генерирует при запуске метода классификации.

Меню «Дополнительные параметры» на панели «Классификация» можно использовать для настройки вывода. В зависимости от настроек Weka создаст один или несколько следующих разделов

Какой из разделов присутствует всегда?

Какой из разделов присутствует при использовании отдельного набора тестов?

В каком разделе используется параметр Folds ?

Теперь давайте более подробно рассмотрим параметры, доступные в диалоговом окне «More options». Какой вариант генерирует код Java, представляющий модель, созданную классификатором (если классификатор предлагает такую возможность)?

Если вы планируете визуализировать прогнозы, сделанные классификатором, какую опцию вам нужно установить?



14. Практическая работа

Откройте набор данных *breast-cancer.arff* в проводнике, перейдите на вкладку Classify и выберите J48.

Одно из значений для *minNumObj* создает то же дерево, что и версия J48 с параметрами по умолчанию (т. е. *unpruned = false*, *minNumObj = 2*). Укажите какой это параметр.

В общем, параметр *trustFactor* в J48 *лучше* не трогать. Но интересно посмотреть на его эффект. Со значениями по умолчанию для других параметров поэкспериментируйте со следующими значениями *trustFactor*, записывая производительность в каждом случае (оценивается с использованием 10-кратной перекрестной проверки): 0.005, 0.05, 0.25, 0.5

Какое значение или значения обеспечивают наибольшую точность?

Лабораторная работа 6

Откройте набор данных *breast-cancer.arff* в проводнике, перейдите на вкладку Classify и выберите J48.

Одним из простых способов сокращения дерева решений является ограничение количества обучающих примеров, достигающих листа. Это делается с помощью параметра `minNumObj` J48 (значение по умолчанию 2) с переключателем `unpruned`, установленным в `True`.

Поэкспериментируйте со следующими значениями `minNumObj`, записывая количество листьев и размер дерева в каждом случае: 1,2,3,5,10,20,50,100

Нарисуйте на график количества листьев в дереве (по вертикальной оси) в зависимости от `minNumObj` (по горизонтальной оси).

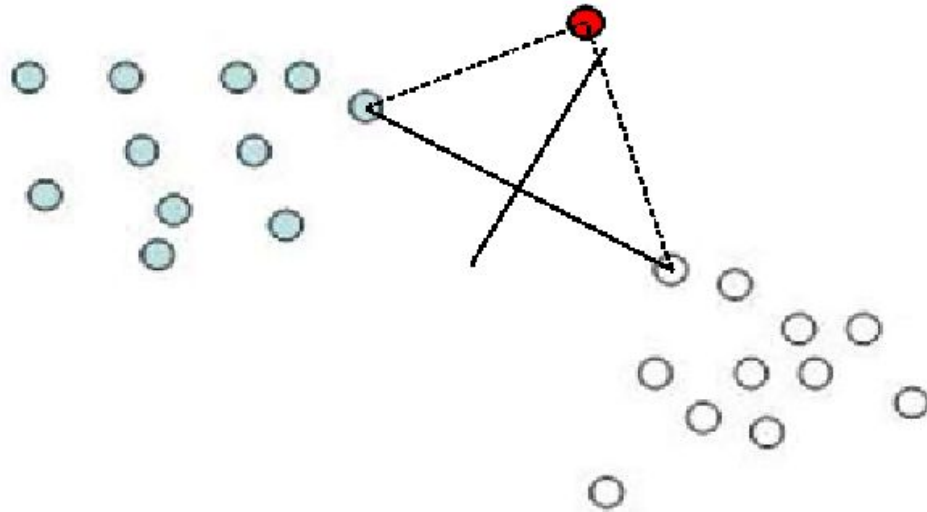
Нарисуйте график при нанесение общего размера дерева (в узлах) на `minNumObj` ?

Ближайший сосед

“Заучивание наизусть”: простейшая форма обучения

- ❖ Чтобы классифицировать новый экземпляр, найдите в тренировочном наборе тот, который “наиболее похож” на него
 - *сами экземпляры представляют “знание”*
 - *ленивое обучение: ничего не делайте, пока вам не нужно делать прогнозы*
- ❖ Обучение “на основе экземпляров” = обучение “ближайший сосед”

Ближайший сосед



Ближайший сосед

Найдите в тренировочном наборе тот, который “наиболее похож” на него

- ❖ Нужна функция подобия
 - Постоянное (“Евклидово”) расстояние? (сумма квадратов разностей)
 - Манхэттенское (“город-квартал (city-block)”) расстояние? (сумма абсолютных разностей)
 - Номинальный атрибуты? Расстояние = 1, если отличается, 0, если одинаково
 - Нормализовать атрибуты, чтобы они лежали между 0 и 1?

Ближайший сосед

Как насчет зашумленных экземпляров?

- ❖ Ближайший сосед (Nearest-neighbor)
- ❖ k -nearest-neighbors
 - *выбрать класс большинства среди нескольких соседей (k из них)*
- ❖ В Weka,
lazy>1Vk (обучение на основе экземпляров)

Ближайший сосед

Исследуйте эффект изменения k

- ❖ Набор данных **Glass**
- ❖ $lazy > IBk, k = 1, 5, 20$
- ❖ 10-кратная перекрестная проверка

$k = 1$	$k = 5$	$k = 20$
70.6%	67.8%	65.4%

Ближайший сосед

- ❖ Часто очень точно... но медленно:
 - сканировать все данные обучения, чтобы сделать каждый прогноз?
 - сложные структуры данных могут сделать это быстрее
- ❖ Предполагается, что все атрибуты одинаково важны
 - Устранение: выбор атрибута или веса
- ❖ Средства защиты от зашумленных экземпляров:
 - Большинство голосует за k ближайших соседей
 - Взвешивание экземпляров в соответствии с точностью предсказания
 - Определить надежные “прототипы” для каждого класса
- ❖ Статистики используют k -NN с 1950-х годов
 - Если размер обучающей выборки $n \rightarrow \infty$ и $k \rightarrow \infty$ и $k/n \rightarrow 0$, ошибка приближается к минимуму

Границы классификации

Визуализатор границ Weka для OneR

- ❖ Откройте **iris.2D.arff**, набор 2D-данных
 - (можно создать его самостоятельно, удалив атрибуты *sepalength* и *sepalwidth*)
- ❖ Выбор графического интерфейса (GUI) Weka:
Visualization>BoundaryVisualizer
 - открыть *iris.2D.arff*
 - *Примечание: длина лепестка (petallength) по X, ширина лепестка (petalwidth) по Y*
 - выберите *rules>OneR*
 - проверить *Данные обучения графика (Plot training data)*
 - нажмите *Start*
 - в *Explorer* изучить правило *OneR*

Границы классификации

Визуализируйте границы для других схем

- ❖ Выберите **lazy>IBk**
 - *Plot training data; нажмите Start*
 - *$k = 5, 20$; обратите внимание на смешанные цвета*
- ❖ Выберите **bayes>NaiveBayes**
 - *Установить для `useSupervisedDiscretization` значение `true`*
- ❖ Выберите **trees>J48**
 - *Связать график с выводом Explorer*
 - *поэкспериментируйте с `minNumbObj = 5 and 10`: контролирует размер листа*

Границы классификации

- ❖ Классификаторы создают границы в пространстве экземпляров
- ❖ Разные классификаторы имеют разную предвзятость
- ❖ Посмотрели OneR, IBk, NaiveBayes, J48
- ❖ Визуализация ограничена числовыми атрибутами и 2D-графиками



15. Практическая работа

Откройте набор данных *breast-cancer.arff* и перейдите на вкладку Классифицировать. Выберите классификатор 1Vk.

- Какая его точность, оцениваемая с помощью 10-кратной перекрестной проверки?

1Vk в KNN параметр определяет число ближайших соседей использования при классификации экземпляра теста, и результат определяется большинством голосов. Значение по умолчанию - 1.

Оцените производительность KNN с 2, 3 и 5 ближайшими соседями. Какие точности вы получаете и почему?

Как вы думаете, эти различия значительны?



15. Практическая работа

Подтвердите свой ответ, запустив IVk со значением по умолчанию 1 для KNN, используя следующие начальные числа случайных чисел, : 1,2,3,4,5. Требуется скрин.

Очевидная проблема с IVk заключается в том, как выбрать подходящее значение для количества используемых ближайших соседей. Если он слишком мал, метод подвержен помехам в данных. Если он слишком велик, решение размывается, покрывая слишком большую площадь пространства экземпляра.

В реализации Weka IVk есть опция, которая может помочь автоматически выбрать лучшее значение. Проверьте информацию о кнопках в «Подробнее» , укажите какая это кнопка.



15. Практическая работа

Давайте искусственно добавим шум в набор данных, определим наилучшее значение для KNN, используя только что обнаруженный вами вариант, и посмотрим, как оно изменяется с уровнем шума.

Откройте набор данных *glass.arff*. Выберите фильтр неконтролируемых атрибутов *addNoise*. Обратите внимание на его панель конфигурации, что по умолчанию он добавляет 10% шума к последнему атрибуту (классу).

Измените это значение на 30% и примените фильтр. На панели «Классификация» выберите *IVk* и настройте его для автоматического определения наилучшего количества соседей. На первый взгляд, параметр KNN теперь избыточен, но на самом деле это не так.

Выясните, что он делает, поэкспериментируя со значениями 1, 10, 20 и проверив, сколько соседей используется. Когда вы запускаете *IVk*, эта информация появляется в разделе выходных данных модели классификатора.



15. Практическая работа

- Какое количество соседей является наилучшим (по определению Weka), когда количество добавленного шума составляет 0%, 10%, 20% и 30%?
- Укажите 4 числа. Не забудьте Undo эффект addNoise фильтра (или перезагрузить набор данных) после каждого эксперимента.



15. Практическая работа

Выберите классификатор IVk с параметрами по умолчанию и запустите визуализацию границ. Вы заметите небольшую слабую область смешанного цвета (зеленого и синего).

- Как можно смешивать цвета, когда используется только один ближайший сосед? Изучите это с помощью панели Визуализация и обоснуйте свой ответ с доказательствами из Weka .



16. Практическая работа

Откройте набор данных *glass.arff*, перейдите на вкладку Classify и используйте процентное разделение со значением по умолчанию 66% в качестве метода оценки.

- Какая точность ZeroR (в процентах)?
- Какая точность J48 в наборе данных о стекле с использованием значений параметров по умолчанию?
- Какая точность NaiveBayes в наборе данных о стекле с использованием значений параметров по умолчанию?

Откройте набор данных *segment-challenge.arff*, перейдите на вкладку Classify.

- Какая точность ZeroR?
- Какая точность IBk для *segment-challenge.arff*, оцениваемого при тестировании сегмента с использованием значений параметров по умолчанию?
- Какая точность PART для *segment-challenge.arff*, оцениваемой при тестировании сегмента с использованием значений параметров по умолчанию?

Лабораторная работа 7

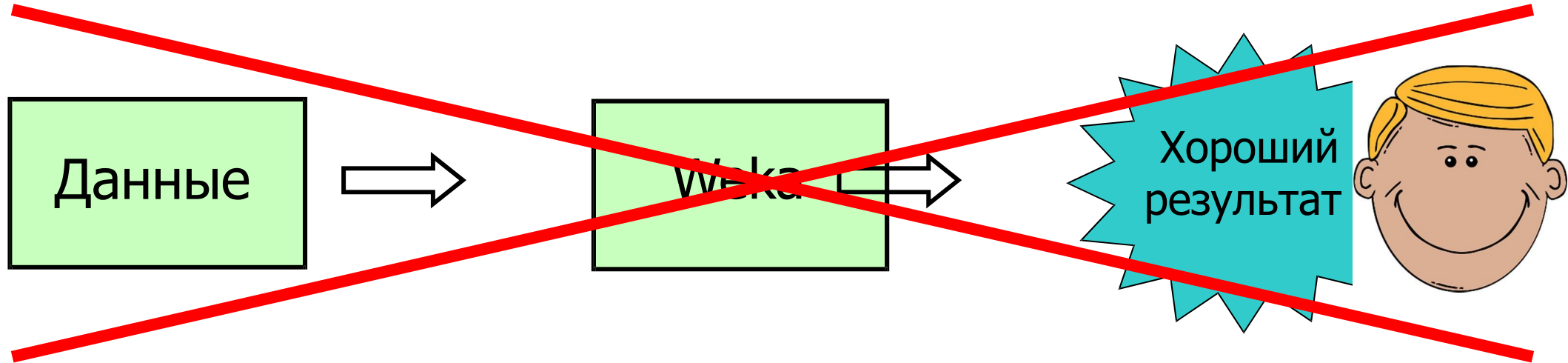
С помощью перекрестной проверки Weka создает модель для каждого разделения.

Какой из них используется для классификации свежих данных, на примере 10-кратной перекрестной проверки? Подсказка разделений 11.

Рискованно ли использовать Weka на практике, если точно не знать, как работают классификаторы?

Главный вопрос недели: «Как работают простые методы классификации? Как работает каждый из них?», на примере рассказа брату, партнеру, родителям.

Процесс интеллектуального анализа данных



Процесс интеллектуального анализа данных



Процесс интеллектуального анализа данных

- ❖ Задайте вопрос
 - *Что вы хотите узнать?*
 - *"Расскажите мне что-нибудь интересное о данных" этого недостаточно!*
- ❖ Соберите данные
 - *вокруг так много всего...*
 - *... но ... нам нужны (экспертные?) классификации*
 - *больше данных побеждает умный алгоритм*
- ❖ Почистите данные
 - *Реальные данные очень грязные*
- ❖ Определите новые функции
 - *разработка функций—ключ к интеллектуальному анализу данных*
- ❖ Раскройте результат
 - *техническая реализация*
 - *Убедите своего босса!*

Процесс интеллектуального анализа данных

(Выбранные) фильтры для разработки функций

❖ AddExpression (MathExpression)

Применение математического выражения к существующим атрибутам для создания новых (или изменения существующих).

❖ Center (Нормализация) (Стандартизация)

– Преобразование числовых атрибутов для получения нулевого значения (или в заданном числовом диапазоне) (или получения нулевого значения и единичной дисперсии)

❖ Discretize (Также контролируемая дискретизация)

– Дискретизация числовых атрибутов для получения номинальных значений

❖ PrincipalComponents

– Выполнение анализа основных компонентов/преобразования данных

❖ RemoveUseless

– Удаление атрибутов, которые совсем не меняются или меняются слишком сильно.

❖ TimeSeriesDelta, TimeSeriesTranslate

– Замена значений атрибутов с различиями между текущим экземпляром и следующим.

Процесс интеллектуального анализа данных

- ❖ Века лишь малая часть (к сожалению) ...
- ❖ ... и это легкая часть

“Пусть все ваши проблемы будут техническими”

– Благословение пожилого программиста

Подводные камни и ловушки

Будьте осторожны

- ❖ Очень легко просчитаться в интеллектуальном анализе данных – *сознательно или бессознательно*
- ❖ Для надежных тестов используйте совершенно новую выборку данных, которую никогда раньше не использовали.

Переобучение очень многогранно

- ❖ Не тестируйте на обучающем наборе (Само собой!)
- ❖ Данные, которые использовались для обучения (любым образом) - портятся.
- ❖ Оставьте некоторые оценочные данные на самый конец.

Подводные камни и ловушки

Отсутствующие значения

“Отсутствующие” значит ...

- ❖ Известные?
- ❖ Незаписанные?
- ❖ Неуместные?

Вы должны: 1. Пропустить случаи, когда значение атрибута отсутствует?
или 2. Рассматривать «отсутствует» как отдельное возможное значение?

Имеет ли значение тот факт, что значение отсутствует?

Большинство алгоритмов обучения работают с пропущенными значениями.

– но они могут делать разные предположения о них.

Подводные камни и ловушки

OneR и J48 работают с пропущенными значениями по разному

- ❖ Запустите `weather-nominal.arff`
- ❖ OneR получает 43%, J48 получает 50% (используя 10-кратную перекрестную проверку)
- ❖ Измените значение `прогнозов` на `unknown` для четырех первых `неопределенных` экземпляров
- ❖ OneR получает 93%, J48 все еще получает 50%
- ❖ Посмотрите на правило OneR: оно использует "?" как четвертое значение в `прогнозе`.

Подводные камни и ловушки

Бесплатных обедов не бывает



- ❖ Задача 2-го класса со 100 бинарными атрибутами
- ❖ Скажем, вы знаете миллион экземпляров и их классы (тренировочный набор).
- ❖ Вы не знаете классов от $2^{100} - 10^6$ примеров
(это 99.9999...% от набора данных)
- ❖ Как вы сможете их понять?

В общем для обобщения, каждый учащийся должен воплотить некоторые знания или предположения, выходящие за рамки данных, которые ему предоставлены.

Алгоритм обучения неявно предоставляет набор предположений. Не может быть «универсального» лучшего алгоритма(бесплатного обеда не бывает).

Интеллектуальный анализ данных - экспериментальная наука

Подводные камни и ловушки

- ❖ Будьте осторожны
- ❖ Переобучение очень многогранно
- ❖ Отсутствующие значения – разные предположения
- ❖ Нет «универсального» лучшего алгоритма обучения
- ❖ Интеллектуальный анализ данных - экспериментальная наука
- ❖ Очень легко просчитаться

Интеллектуальный анализ данных и этика

Законы о конфиденциальности информации (в Европе, но не в США)

- ❖ Для сбора любой личной информации требуется указать цель
- ❖ Такая информация не должна разглашаться другим лицам без согласия
- ❖ Записи о физ. лицах должны быть точными и актуальными
- ❖ Для обеспечения точности люди должны иметь возможность просматривать данные о себе
- ❖ Данные должны быть удалены, когда они больше не нужны для заявленной цели
- ❖ Личная информация не должна передаваться в места, где защита данных не может быть обеспечена должным образом
- ❖ Некоторые данные слишком конфиденциальны, чтобы их можно было собирать, за исключением крайних обстоятельств (например, сексуальная ориентация, религия).

Интеллектуальный анализ данных и этика

Анонимизация сложнее, чем вы думаете

Когда в середине 1990-х годов Массачусетс опубликовал медицинские данные, в которых резюмировались больничные записи каждого государственного служащего, губернатор публично заверил, что они были анонимными, удалив всю идентифицирующую информацию, такую как имя, адрес и номер социального страхования. Он был удивлен, когда получил по почте свои собственные медицинские карты (включая диагнозы и рецепты).

Техники повторной идентификации. Использование общедоступных записей:

- ❖ 50% Американцев могут быть идентифицированы по городу, дате рождения и полу
- ❖ 85% могут быть идентифицированы, если также указать индекс

База данных фильмов на Netflix: 100 миллионов записей по рейтингу фильмов (1–5)

- ❖ Можно идентифицировать 99% людей в базе данных, если известны оценки по 6 фильмам и примерное время, когда человек их смотрел (\pm неделя)
- ❖ Можно идентифицировать 70% людей, если известны оценки по 2 фильмам и и примерное время, когда человек их смотрел.

Интеллектуальный анализ данных и этика

Цель интеллектуального анализа данных состоит в том, чтобы различать ...

- ❖ кто получает кредит
- ❖ кто получает спецпредложение

Некоторые виды разделения неэтичны и незаконны

- ❖ расовые, половые, религиозные, ...

Но это зависит от контекста

- ❖ Половое разделение обычно незаконно
- ❖ ... за исключением врачей, которые должны учитывать пол
- ... и даже информация, которая кажется безобидной не может быть использована
- ❖ Почтовый индекс связан с расой
- ❖ Членство в определенных организациях связано с полом

Интеллектуальный анализ данных и этика

Корреляция не означает причинно-следственную связь

По мере роста продаж мороженого растет и количество утонувших. Следовательно, употребление мороженого вызывает возможность утонуть???

Интеллектуальный анализ данных выявляет корреляцию, а не причинно-следственную связь

но на самом деле мы хотим предсказать последствия наших действий

Интеллектуальный анализ данных и этика

- ❖ Конфиденциальность личной информации
- ❖ Анонимизация сложнее, чем вы думаете
- ❖ Повторная идентификация по якобы анонимным данным
- ❖ Интеллектуальный анализ данных и дискриминация
- ❖ Корреляция не означает причинно-следственную связь

Итоги курса

- ❖ Интеллектуальный анализ данных - это не волшебство
 - *Это огромное количество различных методов и техник*
- ❖ Не существует единого универсального “Лучшего метода”
 - *Это экспериментальная наука!*
 - *Что лучше всего работает с вашей проблемой?*
- ❖ С Века делать это проще
 - *... может быть слишком просто?*
- ❖ Есть много подводных камней
 - *Вы должны понимать, что делаете!*
- ❖ Сосредоточьтесь на оценке ... и значимости
 - *Алгоритмы различаются по производительности – но существенно ли это?*

Итоги курса

Что мы упустили?

- ❖ Фильтрующие классификаторы

Фильтрация обучающих данных, но не тестовых во время перекрестной проверки.

- ❖ Оценка и классификация с учетом затрат

Оценивайте и минимизируйте затраты, а не количество ошибок

- ❖ Выбор атрибутов

Выберите подмножество для использования при обучении

- ❖ Кластеризация

Узнайте что-нибудь, даже если нет значения класса

- ❖ Правила ассоциации

Найдите ассоциации между атрибутами, когда не указан "класс"

- ❖ Классификация текстов

Обработка текстовых данных в виде слов, символов, n-грамм

- ❖ Weka Experimenter

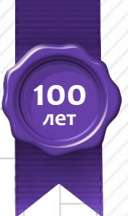
Автоматический расчет средних значений и стандартных отклонений

Итоги курса

- ❖ Данные
 - *Зафиксированные факты*
- ❖ Информация
 - *Шаблоны или предположения, лежащие в их основе*
- ❖ Знания
 - *Накопление вашего набора предположений*
- ❖ Мудрость
 - *Ценность, получаемая со знаниями*

Лабораторная работа 8

С помощью экспериментальной установки «Исследование зрительной системы человека для определения оптимального субъективного качества в потоковом видео МТУСИ» соберите свой собственный набор данных.



**Спасибо
за внимание!**

Контактная информация

111024, г. Москва,
улица Авиамоторная, 8а

АДРЕС ОРГАНИЗАЦИИ

[a.i.mozhaeva@mtuci.](mailto:a.i.mozhaeva@mtuci.ru)
ГУЕ-MAIL