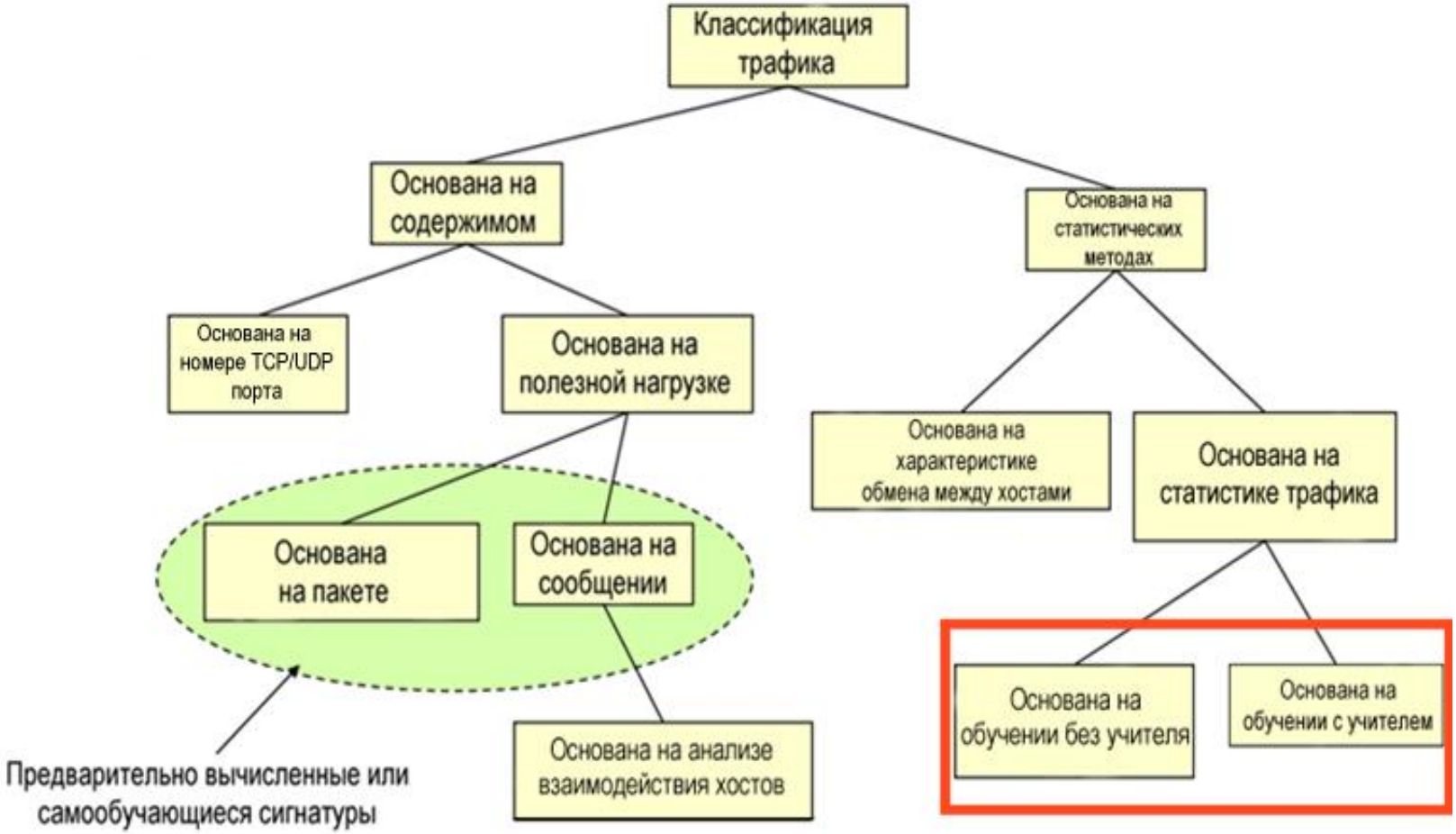
Идентификация сетевых протоколов прикладного уровня на основе методов распознавания образов

Выполнил студент Д.Ю.Федоров

Таксономия способов классификации IP-трафика



Постановка задачи

Формальная постановка задачи:

рассмотрим

$$f : X \to Y,$$

где X — набор векторов атрибутов сетевых пакетов, Y — набор наименований классов. Значения целевой зависимости f известны только на объектах конечной обучающей выборки

$$X^m = \{(x_1, y_1), ..., (x_m, y_m)\}.$$

Требуется построить алгоритм

$$a : X \to Y$$

способный классифицировать произвольный объект $x \in X$.

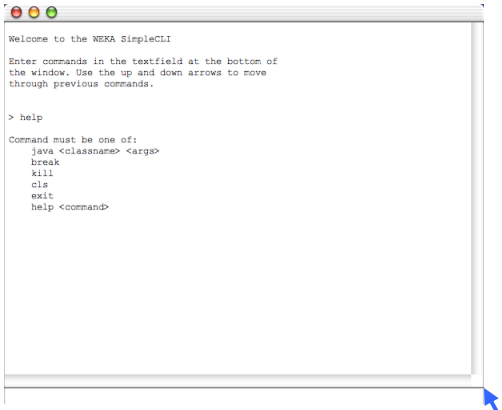
Ограничения на атрибуты:

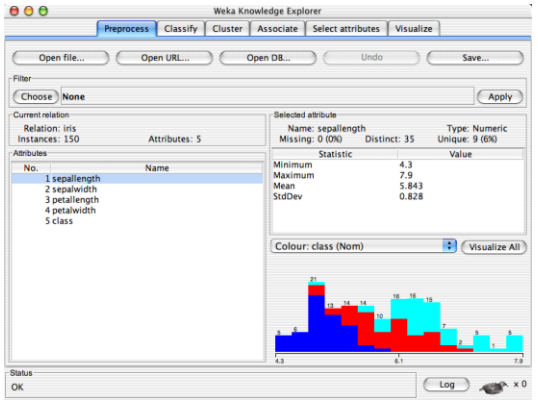
в качестве атрибутов классификации должны использоваться свойства пакетов транспортного (TCP-протокол) и сетевого (IP-протокол) уровней.

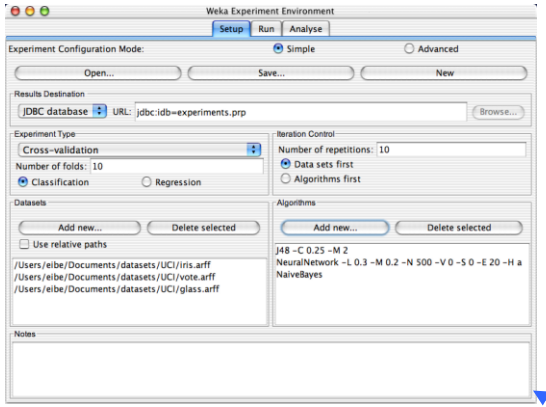
Дополнительно:

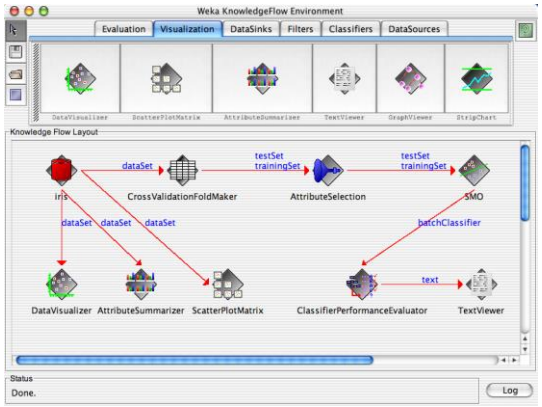
исследовать возможность применения методов сокращения числа признаков и применения методов кластеризации

Библиотека алгоритмов машинного обучения Weka



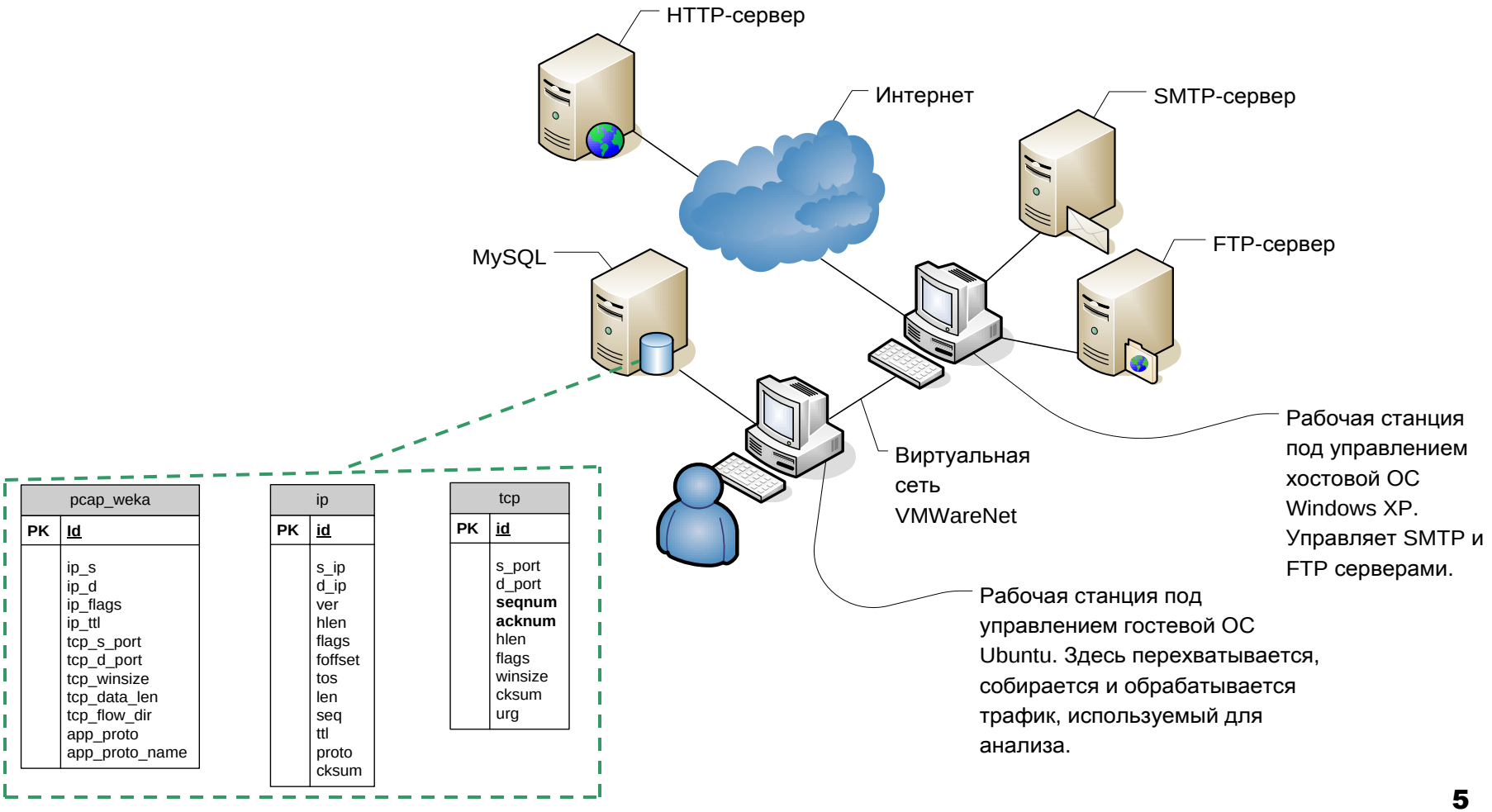






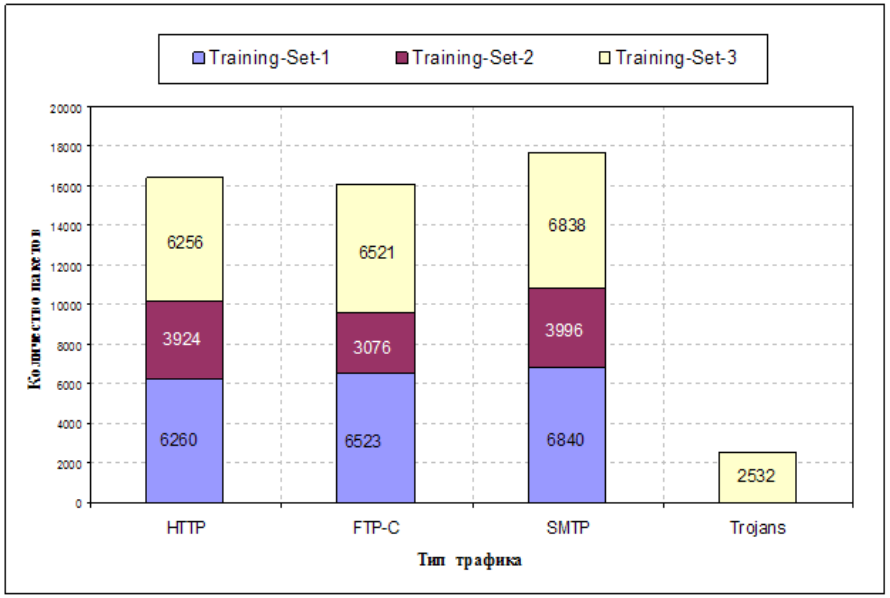
Подготовка данных

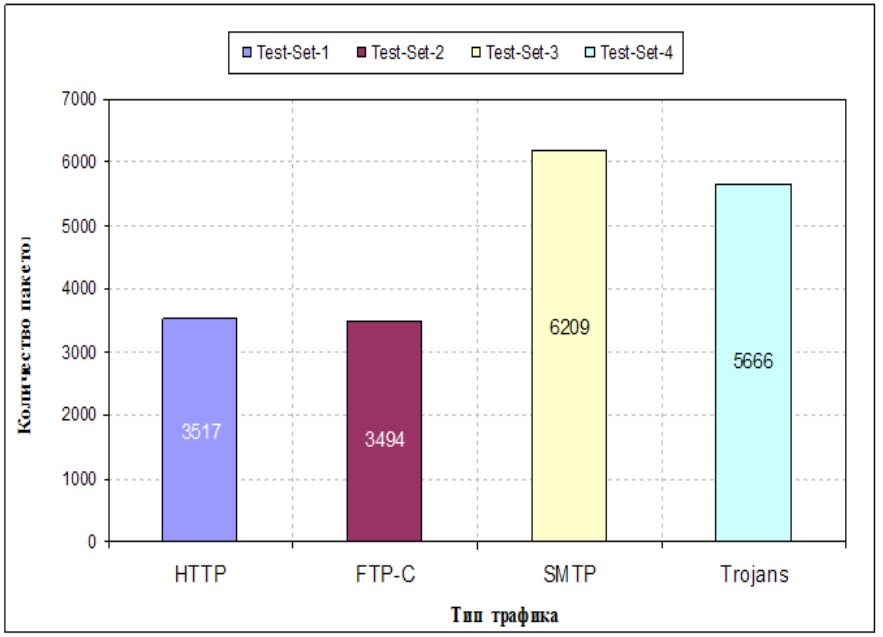
Захват трафика и разбор содержимого пакетов: *Wireshark,* сохранение в формате *tcpdump,* утилита *pcap2mysql*



Подготовка данных

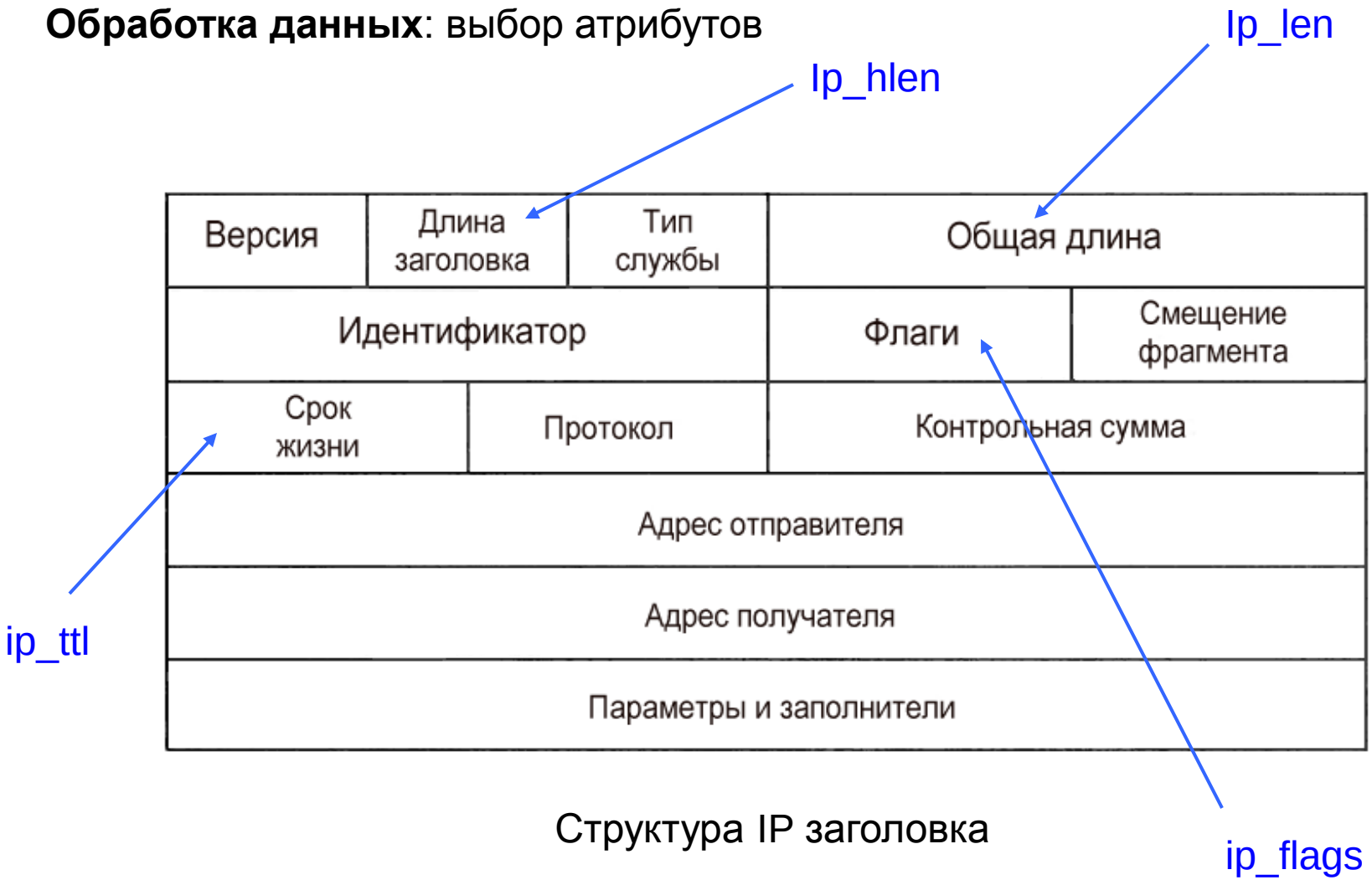
Обработка данных: предварительная классификация трафика с помощью PHP-скрипта на основании известного номера TCP-порта





Подготовка данных

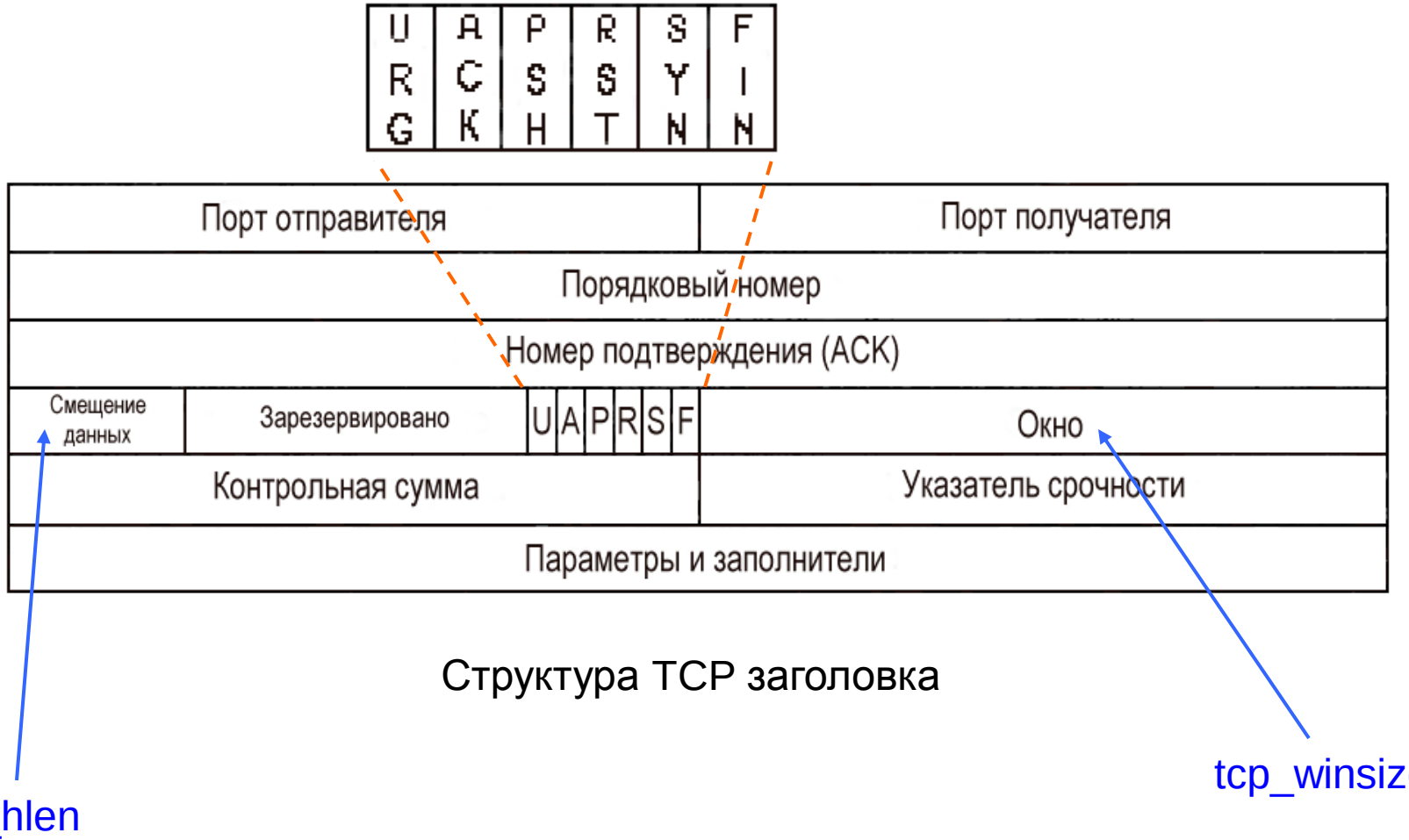
Обработка данных: выбор атрибутов



Структура IP заголовка

Подготовка данных

Обработка данных: выбор атрибутов



Структура TCP заголовка

Подготовка данных

Обработка данных: выбранные атрибуты

- IP-флаги (ip_flags)
- Срок жизни (ip_ttl)
- Окно (tcp_winsize)
- Размер полезной нагрузки (tcp_data_len), вычисляется по формуле: ip_len-ip_hlen*4-tcp_hlen*4
- Направление потока, от сервера к клиенту/от клиента к серверу (tcp_flow_dir)

Методы выделения признаков

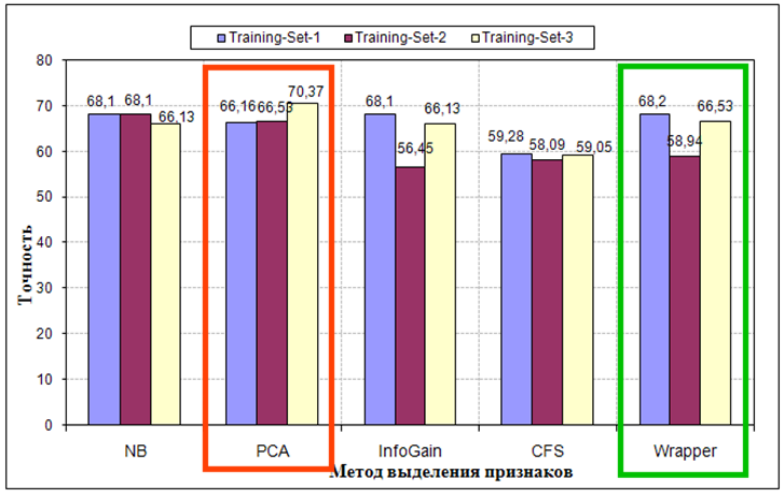
- Выбраны методы:
  - PCA
  - InfoGain
  - CFS
  - Wrapper

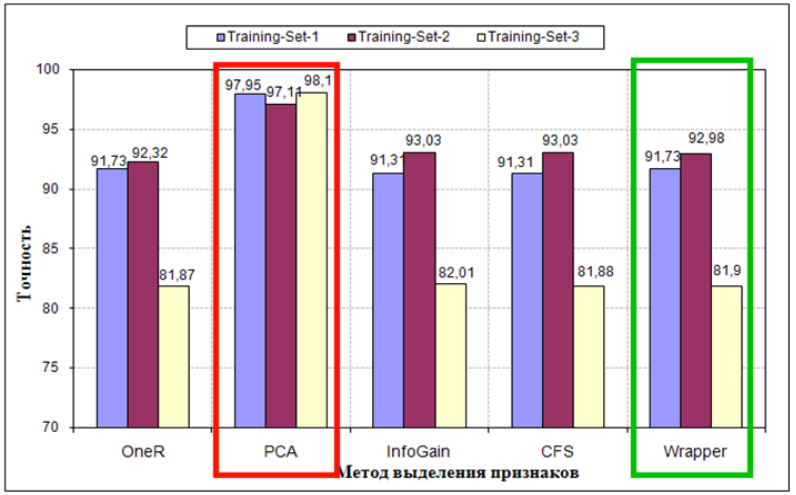
- Задача эксперимента:
  - определить какие признаки были выделены;
  - определить точность классификации до и после выделения признаков.

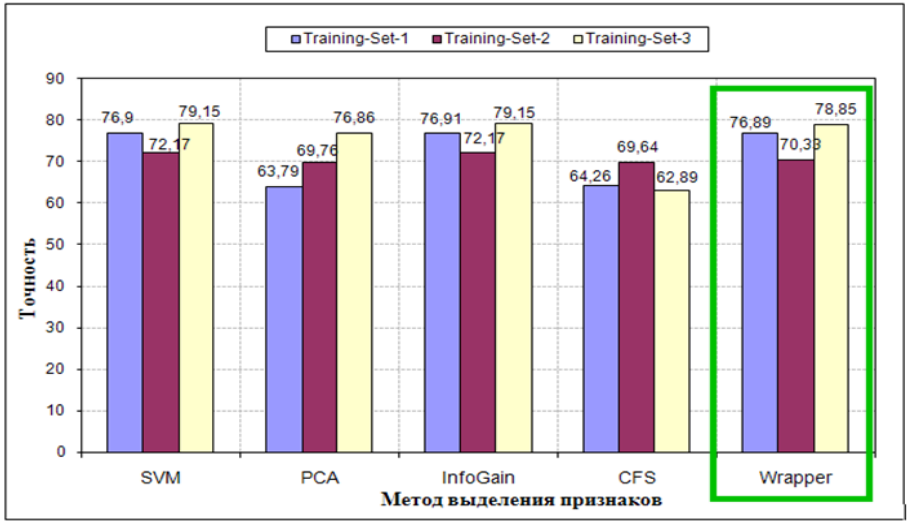
Какие признаки были выделены?

Название метода	Выбранные атрибуты	Комментарии
PCA	-0.506 ip_flags + 0.495 ip_ttl + 0.491 tcp_flow_dir + 0.471 tcp_winsize + 0.192 tcp_data_len 0.927 tcp_data_len - 0.264 ip_ttl + 0.217 ip_flags + 0.15 tcp_flow_dir - 0.023 tcp_winsize - 0.855 tcp_winsize - 0.402 ip_flags + 0.288 ip_ttl + 0.146 tcp_data_len + 0.059 tcp_flow_dir	Выбраны 3 признака, состоящие из линейно комбинации элементов исходных признаков с коэффициентами в виде собственных векторов
InfoGain	tcp_data_len tcp_winsize ip_ttl ip_flags tcp_flow_dir	Результатом работы метода является список признаков, ранжированных по их значимости
CFS	tcp_winsize tcp_data_len	
Wrapper	ip_flags tcp_winsize tcp_data_len tcp_flow_dir	Результаты метода Wrapper в отличие от остальных рассмотренных методов, зависят от индукционного алгоритма

Точность классификации до и после выделения признаков







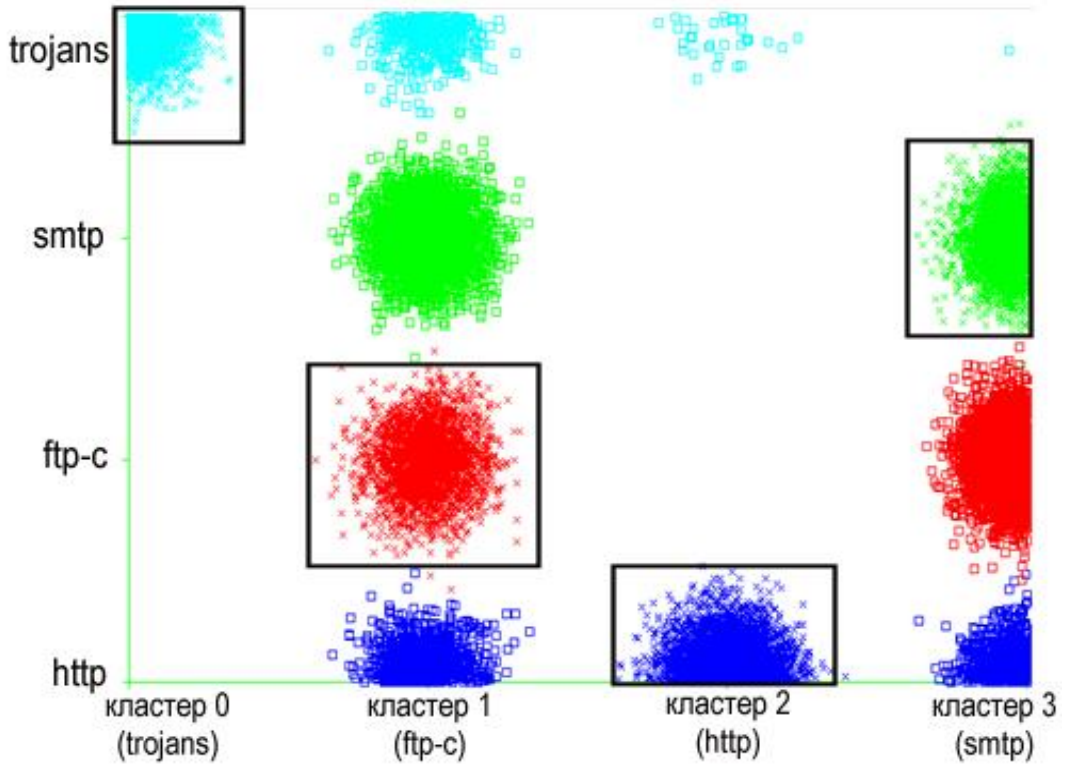
Методы кластеризации

- Выбраны методы:
  - EM
  - k-средних

- Задача эксперимента:

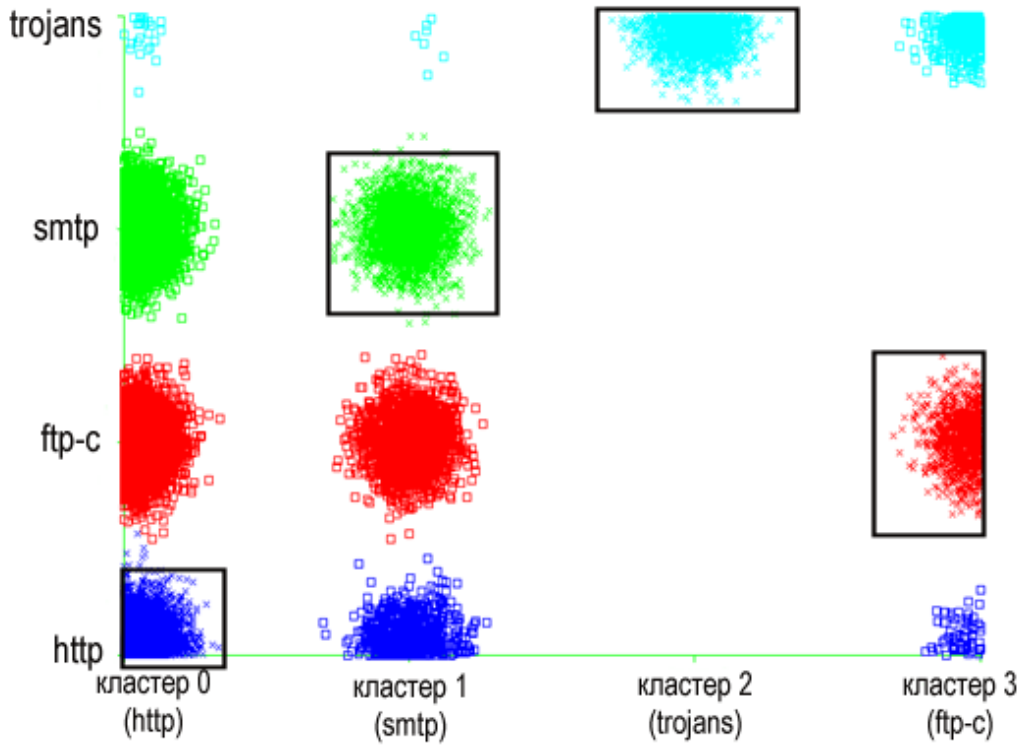
  сопоставить полученные кластеры с предварительно классифицированными сетевыми пакетами

Результат кластеризации с помощью метода EM



Процент ошибочно кластеризованных пакетов – 38.9%.

Результат кластеризации с помощью метода k-средних



Процент ошибочно кластеризованных пакетов – 46.8%.

Методы классификации

- Выбраны методы:
  - Naïve Bayes
  - J4.8
  - SVM
  - OneR

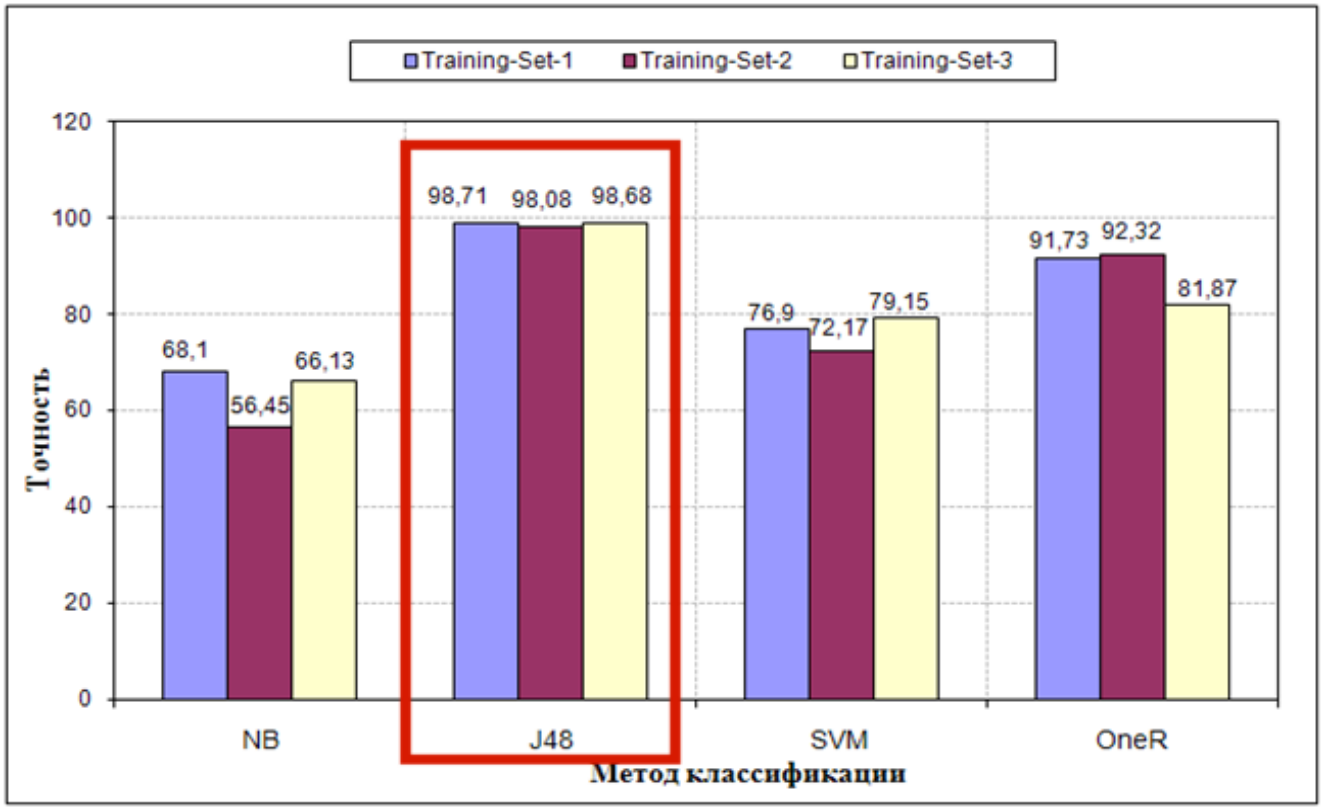
- Задача эксперимента:

  сравнить результаты методов классификации для тестовых и обучающих множеств

Сравнение методов классификации для тестовых множеств (процент ошибочно классифицированных пакетов)

Название дампа	Naïve Bayes	J4.8	SVM	OneR	Итог
Test-Set-1	21.3	3	6	12.4	J4.8 (HTTP)
Test-Set-2	28.2	2	61.4	5.3	J4.8 (FTP-C)
Test-Set-3	64	1	0	22.7	J4.8, SVM (SMTP)
Test-Set-4	12	7.7	2.7	90.6	SVM (Trojans)

Сравнение методов классификации для обучающих множеств



Результаты и перспективные направления исследования

Выполнено:

- классификация сетевого трафика и сравнение точности выбранных методов классификации;
- выделение признаков и сравнение точности классификации до и после применения методов выделения признаков;
- кластеризация и сопоставление кластеризованных сетевых пакетов с заранее классифицированными.

Результаты:

- наилучшие результаты при классификации показал метод J4.8 (точность 98,71%);
- точность классификации увеличилась в результате применения метода выделения признаков PCA;
- алгоритмы кластеризации EM и k-средних показали плохие результаты (38,9% и 46,8% ошибочно кластеризованных пакетов соответственно).

Перспективные направления исследования:

- идентификация трафика в реальном времени;
- захват потоков, вместо отдельных пакетов;
- увеличение числа атрибутов (учет производных аргументов).