

Оглавление

Введение.....	3
1. КЛАССИФИКАЦИЯ. ОСНОВНЫЕ ПОНЯТИЯ, ЗАДАЧИ И ПРОБЛЕМЫ	6
1.1. Задачи классификации IP-трафика.....	6
1.2. Методы классификации сетевого трафика.....	8
1.2.1. Классификация IP-трафика на основе портов.....	9
1.2.2. Классификация сетевого трафика на основе полезной нагрузки.....	11
1.2.3. Классификация на основе статистических методов.....	15
1.2.4. Особенности применения методов машинного обучения для классификации сетевого трафика.....	18
1.2.5. Статистическая кластеризация.....	23
1.2.6. Иные подходы.....	24
Литература.....	25
2. КЛАССИЧЕСКИЕ ПАРАДИГМЫ МАШИННОГО ОБУЧЕНИЯ И ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА ДАННЫХ	33
2.1. Основные понятия. Технологии KDD и Data Mining.....	33
2.2. Классификация. Основы обучения с учителем.....	39
2.2.1. Классификация на основе ассоциативных правил.....	40
2.2.2. Искусственные нейронные сети (ИНС).....	41
2.2.3. Метод опорных векторов.....	43
2.2.4. Решающие деревья.....	45
2.2.5. Алгоритм ID3.....	54
2.2.6. Алгоритм C4.5.....	54
2.2.7. Алгоритм CART.....	57
2.2.8. Алгоритм CHAID.....	59
2.2.9. Алгоритм QUEST.....	59
2.2.10. Алгоритм случайного леса.....	60
2.2.11. Алгоритмы Bootstrap, Bagging и AdaBoost.....	62
2.2.12. Наивный байесовский классификатор.....	65
2.2.13. Байесовские сети.....	66
2.2.14. Оценка устойчивости классификатора.....	66
2.2.15. Методы поиска аномалий, основанные на классификации.....	68

2.3. Кластеризация. Основы обучения с учителем	69
2.3.1. Основные понятия	69
2.3.2. Методы кластерного анализа данных	71
2.3.3. Иерархические методы	74
2.3.4. Неиерархические методы	81
2.3.5. Неинкрементальные алгоритмы	88
2.3.6. Сравнительный анализ методов кластеризации	92
2.3.7. Самоорганизующаяся карта Кохонена	95
2.3.8. Генетические алгоритмы	98
2.3.9. Достоинства и недостатки методов кластеризации	98
2.3.10. Методы поиска аномалий, основанные на кластеризации	100
2.4. Метрики оценки эффективности классификации и кластеризации	102
2.5. Инструменты для интеллектуального анализа данных	106
2.5.1. Rattle	106
2.5.2. Weka	106
2.5.3. MOA	108
2.5.4. Orange	111
2.5.5. RapidMiner	112
2.5.6. Scikitlearn	112
2.6. Проблемы машинного обучения (контролируемое и неконтролируемое обучение)	113
Литература	114
3. АНАЛИЗ И МОНИТОРИНГ СЕТЕВОГО ТРАФИКА	117
3.1. Проблемы контроля и анализа сетевого трафика	117
3.1.1. Место контроля трафика	118
3.1.2. Задачи контроля	119
3.2. Сетевые анализаторы трафика	122
3.2.1. Задачи анализа сетевого трафика	122
3.2.2. Средства анализа сетевого трафика	123
3.2.3. Программный сниффер Wireshark	126
3.2.4. Аппаратный сниффер network associates	128
3.2.5. Iris Network Traffic Analyzer	129
3.3. Сбор данных с помощью протокола NetFlow	129
3.3.1. Мониторинг	129
3.3.2. Примеры контрольных и аналитических инструментов потока сетевого трафика с помощью протокола NetFlow	130
3.4. Сбор данных с помощью протокола SNMP	133

3.4.1. Контроль сетевых устройств	133
3.4.2. Примеры контрольных и аналитических инструментов потока сетевого трафика помощью протокола SNMP	134
3.5. Программный сниффер Tcprdump	136
3.6. Другие технологии и подходы к сетевому мониторингу	138
3.6.1. Трассировка событий сетевого стека	138
3.6.2. Протокол ICMP	139
3.6.3. Анализ системных журналов	139
3.7. Инструменты классификации. Технология DPI	140
3.7.1. PACE	141
3.7.2. OpenDPI	141
3.7.3. nDPI	142
3.7.4. Libprotoident	142
3.7.5. Cisco NBAR	142
3.7.6. L7-фильтр	143
3.8. Использование инструментов DPI для классификации и учета трафика	143
3.8.1. Использование инструментов DPI для классификации трафика	143
3.8.2. Использование инструментов DPI для целей учета трафика	145
3.8.3. Влияние усечения пакетов и потоков на классификацию трафика	146
Литература	147
4. КЛАССИФИКАЦИЯ ТРАФИКА МЕТОДАМИ МАШИННОГО ОБУЧЕНИЯ	150
4.1. Анализ алгоритмов выбора атрибутов классификации	150
4.2. Формирование исходных данных и анализ программного обеспечения	157
4.2.1. Методы захвата трафика	157
4.2.2. Результаты применения программного обеспечения	160
4.2.3. Выбор атрибутов классификации	162
4.3. Влияние структуры обучающей выборки на эффективность классификации приложений	165
4.3.1. Процедура сбора трафика	166
4.3.2. Обучающие выборки	167
4.3.3. Выбор атрибутов для классификации	168
4.3.4. Результаты эксперимента	170
4.4. Эффективность алгоритмов выделения атрибутов	173
4.4.1. Формирование исходных данных	173
4.4.2. Сравнительные оценки алгоритмов выделения информативных признаков	174

4.4.3. Результаты классификации	176
4.5. Влияние объема обучающей выборки на качество классификации	179
4.5.1. Алгоритм SVM	179
4.5.2. Алгоритм AdaBoost	180
4.5.3. Классификатор наивный классификатор Байеса ...	181
4.5.4. Алгоритм CART	181
4.5.5. Случайный лес	182
4.6. Эффективность алгоритма RF в задачах классификации приложений	184
4.6.1. Формирование данных	184
4.6.2. Методология решения задачи классификации с помощью алгоритма Random Forest	185
4.6.3. Результаты классификации	188
4.7. Классификация трафика мобильных сетей	191
4.7.1. Захват и анализ сетевого трафика мобильных сетей (приложений)	191
4.7.2. Результаты классификации	195
4.8. Влияние прореживания пакетов на качество классификации	198
4.8.1. Формирование и анализ исходных данных	199
4.8.2. Результаты классификации	200
4.9. Влияние фонового трафика на качество классификации	203
4.9.1. Постановка задачи	204
4.9.2. Результаты классификации	204
4.9.3. Сравнительный ROC — анализ работы алгоритмов при наличии фонового трафика	208
4.10. Классификация шифрованного трафика	210
4.10.1. Формирование и характеристики используемых наборов данных	210
4.10.2. Классификация трафика с помощью формирования сетевых потоков	212
4.10.3. Классификация трафика на основе анализа каждого захваченного сетевого пакета	221
4.11. Интеграция множества классификаторов	226
4.11.1. Ансамбли классификаторов	226
4.11.2. Распространенные типы классификаторов	229
4.11.3. Мультиклассификационная модель классификации	232
4.12. Классификация в режиме реального времени	236
4.12.1. Сценарии обработки потоковых данных	237

4.12.2. Технология смещения концепций	239
4.12.3. Эволюция алгоритмов потоковой классификации	242
4.12.4. Алгоритмы классификации, основанные на потоковых деревьях принятия решений	244
4.12.5. Динамический потоковый Random Forests	249
4.13. Неконтролируемая кластеризация сетевого трафика	254
4.13.1. Технологии кластеризации	254
4.13.2. Кластеризация методом Random Forest и RF близость	257
4.13.3. Кластеризация на основе близости RF	258
4.13.4. Наборы данных	259
4.13.5. Методы оценки	260
4.14. Полуконтролируемая кластеризация сетевого трафика	263
4.14.1. Источники дополнительной информации	263
4.14.2. Алгоритм кластеризации с ограничениями	265
4.14.3. Статистика дополнительной информации	267
4.15. Проблемы классификации трафика	270
Литература	274