Ежегодная международная научно-практическая конференция
«РусКрипто’2020»

Подход к классификации последовательностей, сформированных алгоритмами сжатия и шифрования

Козачок Александр Васильевич,
д.т.н., сотрудник Академии ФСО России, г. Орёл
Спирин Андрей Андреевич,
сотрудник Академии ФСО России, г. Орёл

Анализ объекта исследования

Рисунок 1. Число зарегистрированных утечек информации за 2006 – 2018 гг.

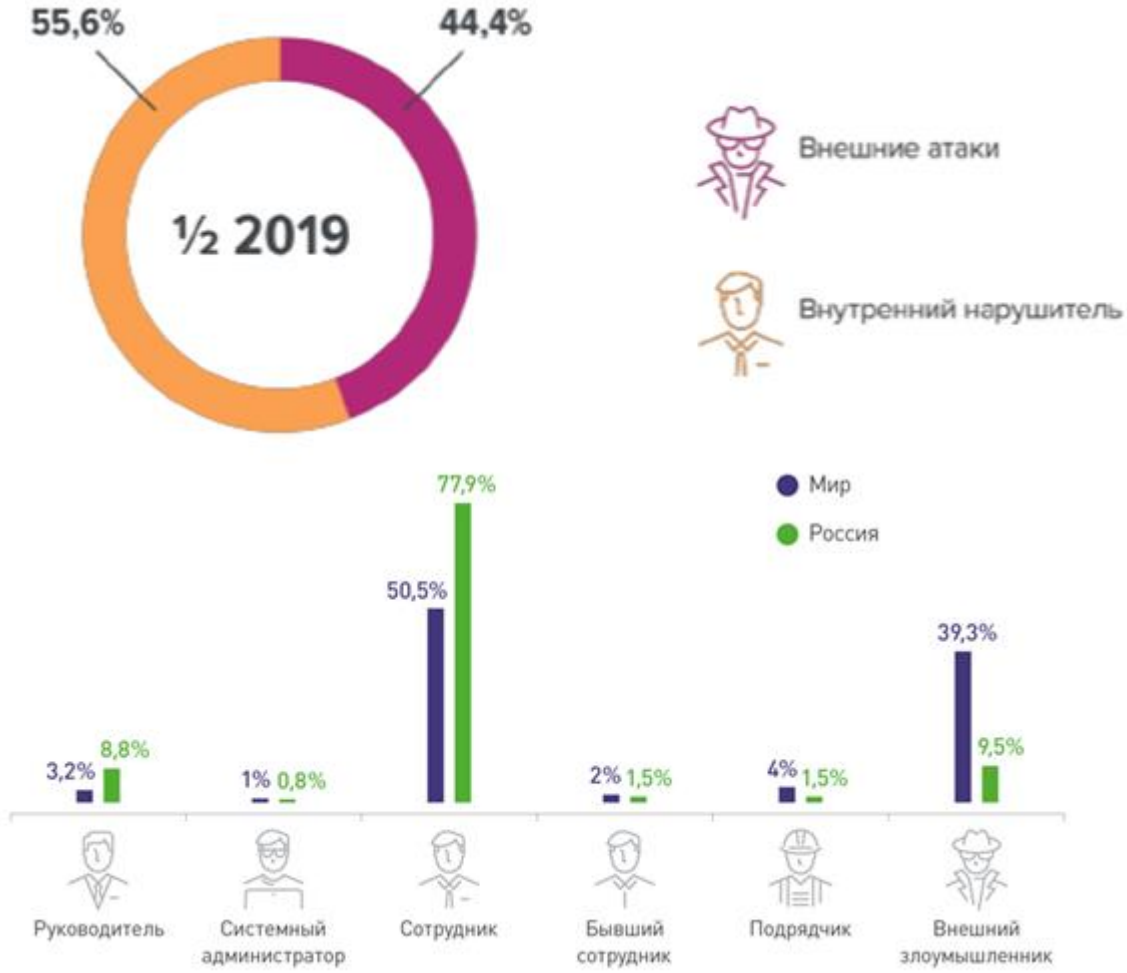


Рисунок 2. Соотношение нарушителей информационной безопасности

*По данным аналитического центра компании INFOWATCH (https://www.infowatch.ru/)

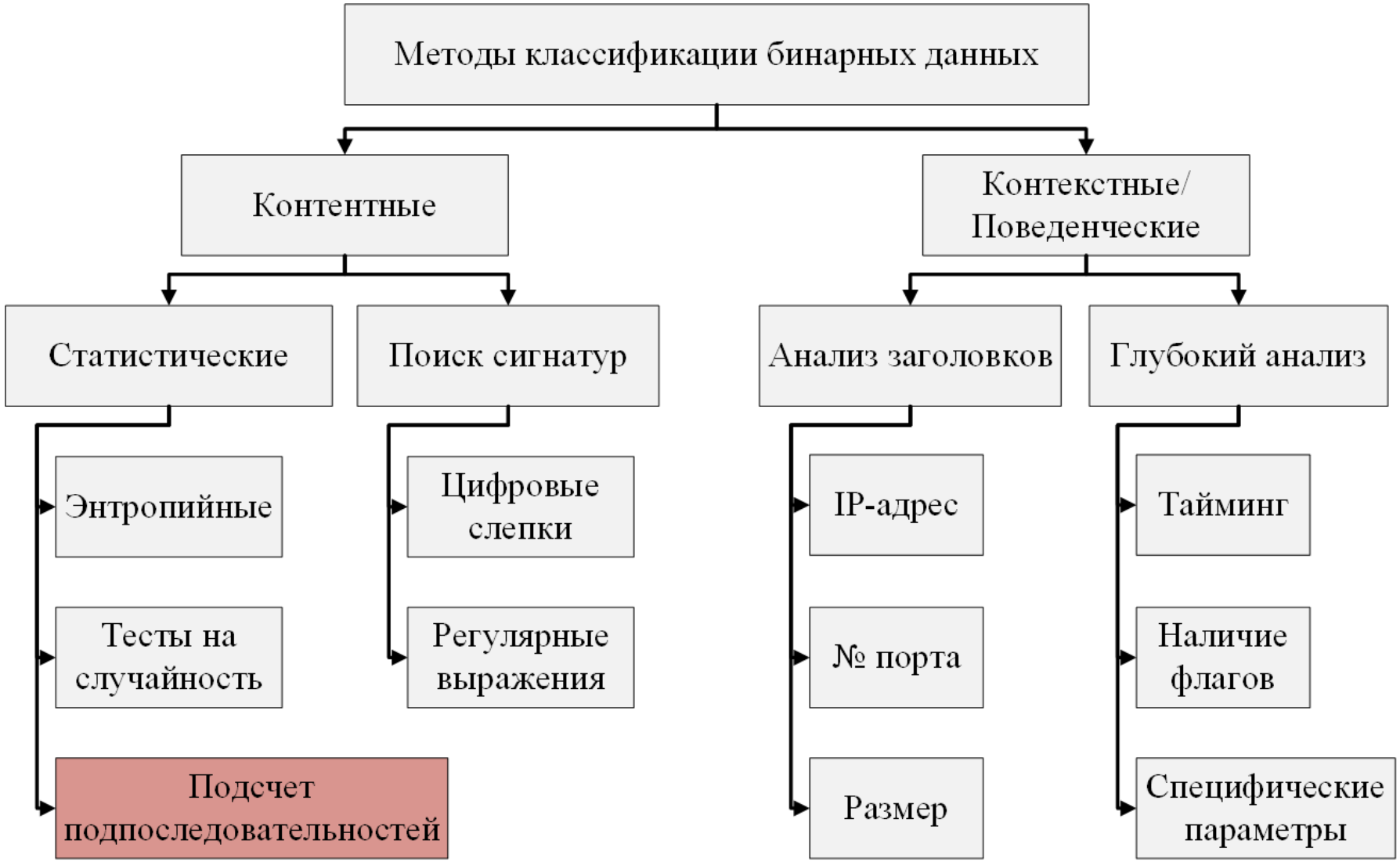
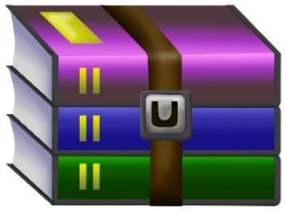


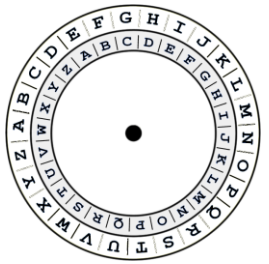
Рисунок 3. Методы, применяемые для классификации шифрованных / сжатых данных

Анализ предмета исследования II

Авторы	Год	Задача	Признаки	Алгоритм	Результаты
Wright C., Monrose F., Masson G. M.	2005	Идентификация сетевых протоколов	Длина пакета, время жизни,	Скрытые Марковские Модели	HTTPS - 94.4%
Conti G.	2010	Классификация бинарных файлов	IP-адреса	(алгоритм Витерби)	HTTP - 96.7%
Alshammari R.	2011	Идентификация зашифрованного трафика	ip.flags, ip.len, ip.checksum, tcp.len*, tcp.flags*, oth.	C4.5, AdaBoost,team-based GA	SSH-71%
Dorfinger P. , Panholzer G., John W.	2011	Идентификация зашифрованного трафика	Оценка энтропии полезной нагрузки	Критерий совпадения	eDonkey, Skype - 94% SMTP, HTTP, POP3, FTP -99%
Wang Y., Zhang Z., Guo L., Li S.	2011	Классификация трафика	Энтропийный подход, подсчет 4-битных подпоследовательностей	SVM	86.4% for text 79.8% enc
Zhang H. , Papadopoulos C., Massey D.	2013	Идентификация зашифрованного трафика botnet	Энтропия потоков и пакетов	Метод Монте-Карло	flow (SSH,HTTPS) -95% packet (SSH,HTTPS) - 97%
Khakpour A. R., Liu A. X.	2013	Идентификация бинарных, текстовых, зашифрованных данных	Энтропия	CART, SVM	88%
Hahn D., Apthorpe N., Feamster N.	2018	Идентификация зашифрованных и сжатых данных	Энтропия, Хи-квадрат	k - ближайших соседей, сверточные нейронные сети, сети прямого распространения	k-Nearest Neighbors 60.0% Feed Forward Neural Net 54.1% CNN 66.9%
Casino F. , Choo K.-K. R., Patsakis C.	2019	Классификация зашифрованного/сжатого трафика	Тесты NIST: frequency block test, cum sums test, approximate entropy test.	HEDGE (High Entropy DistinGuishEr)	70.61%
Tang Z. , Zeng X., Sheng Y.	2019	Классификация зашифрованного/сжатого трафика	Энтропия 4,8,16,24 битных подпоследовательностей в последовательностях, полученных из исходной с различным шагом (от 1 до 8)	SVM, RF	for traffic (ip-адреса, № портов, протокол) 97.9% video <70% Images, text <72% audio <66%

Таблица 1. Сравнительный анализ исследований в предметной области





$$F : X \to Y \quad (1)$$

$$p(F(x_i) = y_j \mid i = j) \to 1 \quad (2)$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (3)$$

$$Precision = \frac{TP}{TP + FP} \quad (4) \qquad Recall = \frac{TP}{TP + FN} \quad (5)$$

где TP – количество объектов верно отнесенных к классу i, TN – количество объектов верно отнесенных к классу j (j≠i), FP – количество ложных срабатываний (ошибка первого рода), FN – количество пропусков цели (ошибка второго рода).

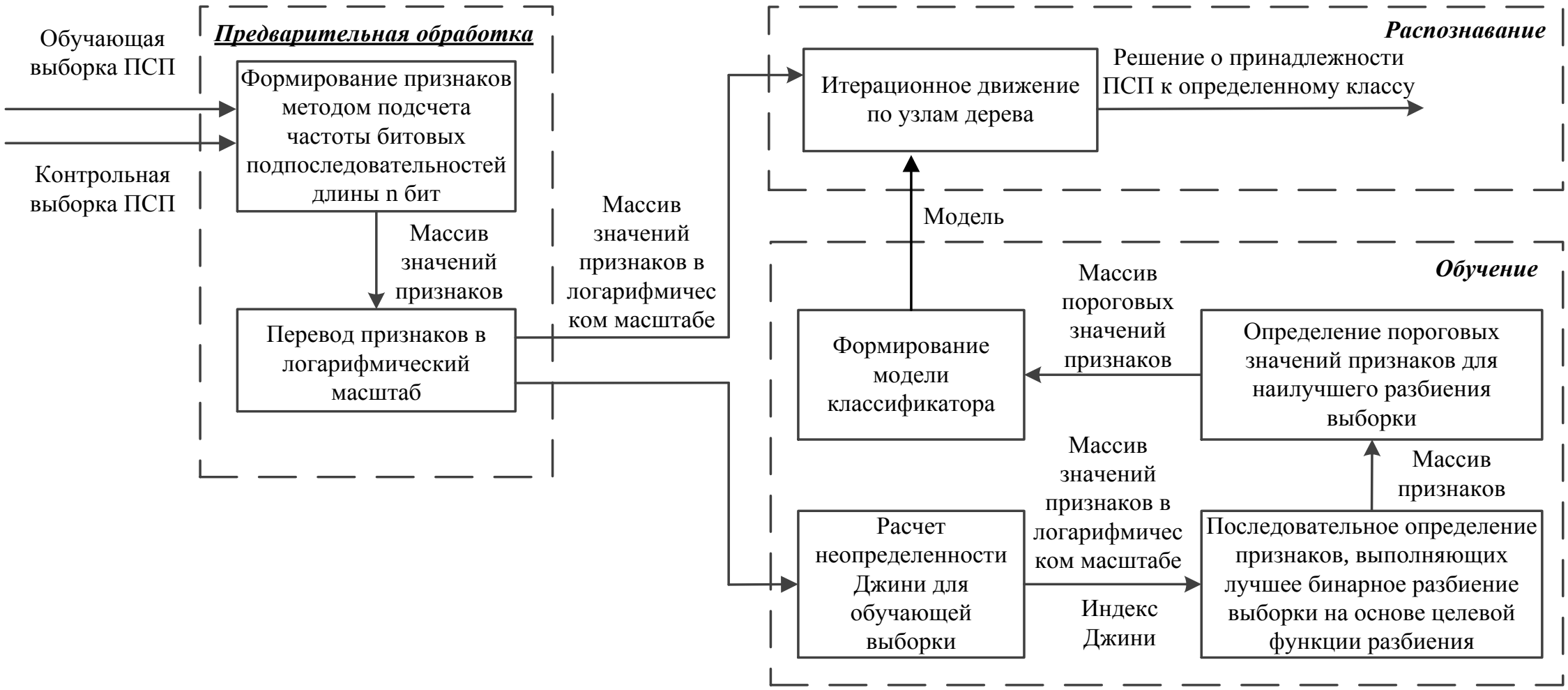


Рисунок 4. Функциональная модель системы классификации последовательностей, сформированных алгоритмами сжатия и шифрования

Data: P: |P|=Q, S: $|S| = 2^{N-1}$
Result: $F_{Q,E}$

$F_{Q,E} \leftarrow <>$
for $p \in P$ **do**
  $M_p \leftarrow$ **Len**(p)
  for $s \in S$ **do**
    $n_s \leftarrow$ **Count**(p,s)
    $f_{p,s} \leftarrow \frac{n_s}{M_p - N_s + 1}$
    $F_{Q,E} \leftarrow F_{Q,E} \cup < f_{p,s}, y_i >$
return $F_{Q,E}$

а)

Data: ПСП p, классификатор <K>, <V>
Result: Класс у ПСП p

$F_{Q,V} \leftarrow <>$
$State \leftarrow <>$
$M_p \leftarrow$ **Len**(p)
for $v \in V$ **do**
  $N_v \leftarrow$ **Len**(v)
  $n_v \leftarrow$ **Count**(p,v)
  $f_{p,v} = \frac{n_v}{M_p - N_v + 1}$
  $F_{Q,V} = F_{Q,V} \cup f_{p,v}$
$State \leftarrow$ **Next** (k)
while $State[7] \neq True$ **do**
  if $f_{p,State[2]} \geq State[3]$ **then**
    $State \leftarrow$ **NextRight** $(State)$
  else
    $State \leftarrow$ **NextLeft** $(State)$
$y_p \leftarrow State[4]$
return y_p

б)

Рисунок 5. Алгоритмы выделения признаков (а) и классификации (б).

Исходные данные

Исходные данные	Алгоритм преобразования	Метка класса	Выборка, кол-во файлов	Размер файла в выборке, Кб
Текст	AES(CBC)	0	2000	600
	3DES(CBC)	0	2000	600
	Camellia(CBC)	0	2000	600
	RC4(CBC)	0	2000	600
	Кузнечик(ECB)	0	2000	600
Видео-файлы	ZIP	1	2000	600
	RAR	1	2000	600
	7Z	1	2000	600
	XZ	1	2000	600
	GZ	1	2000	600
	BZ2	1	2000	600

Таблица 2. Исходные данные для проведения исследования

Результаты экспериментов I

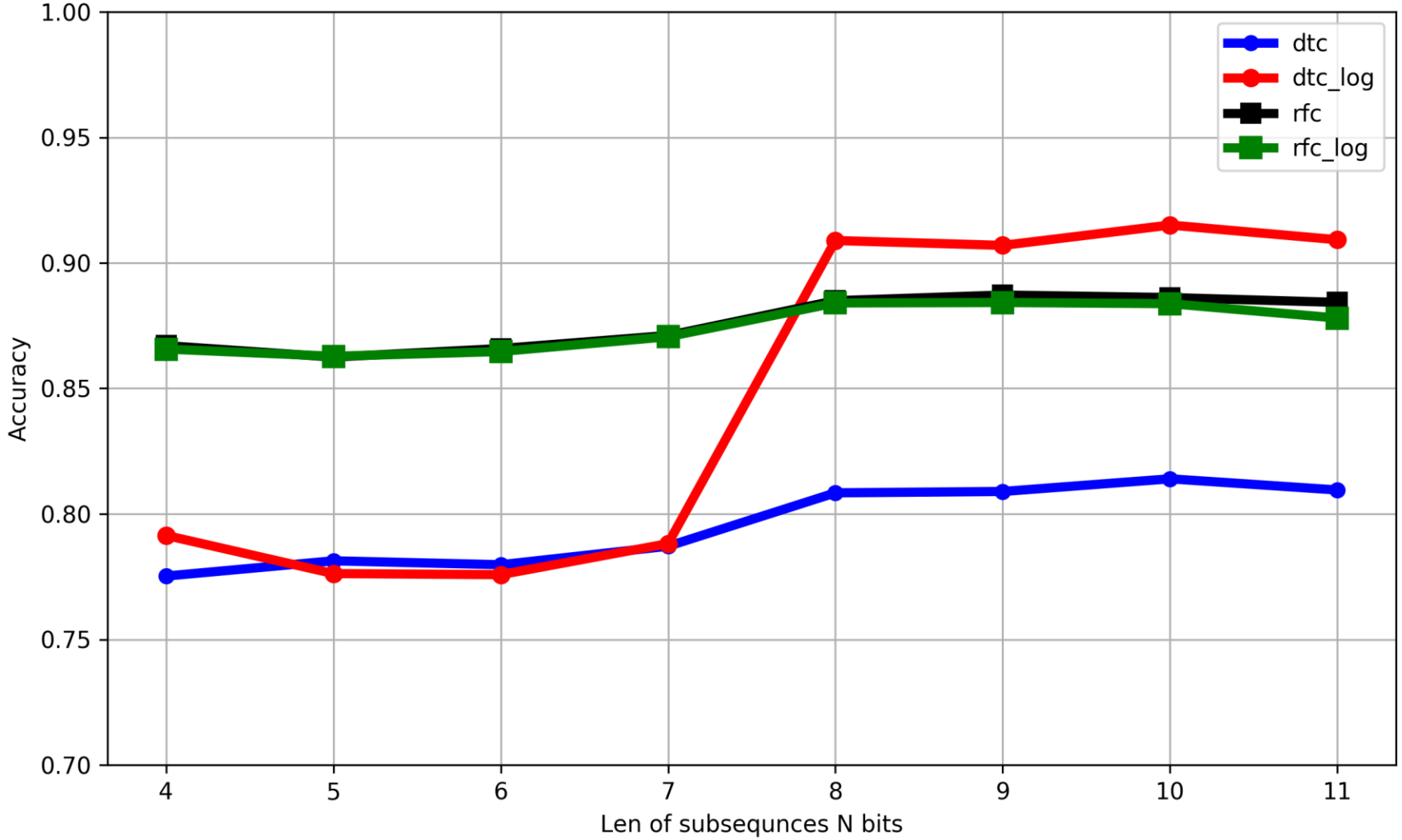


Рисунок 6. Оценка зависимости точности (Accuracy) классификации от длины подпоследовательностей (признаков)

Результаты экспериментов II

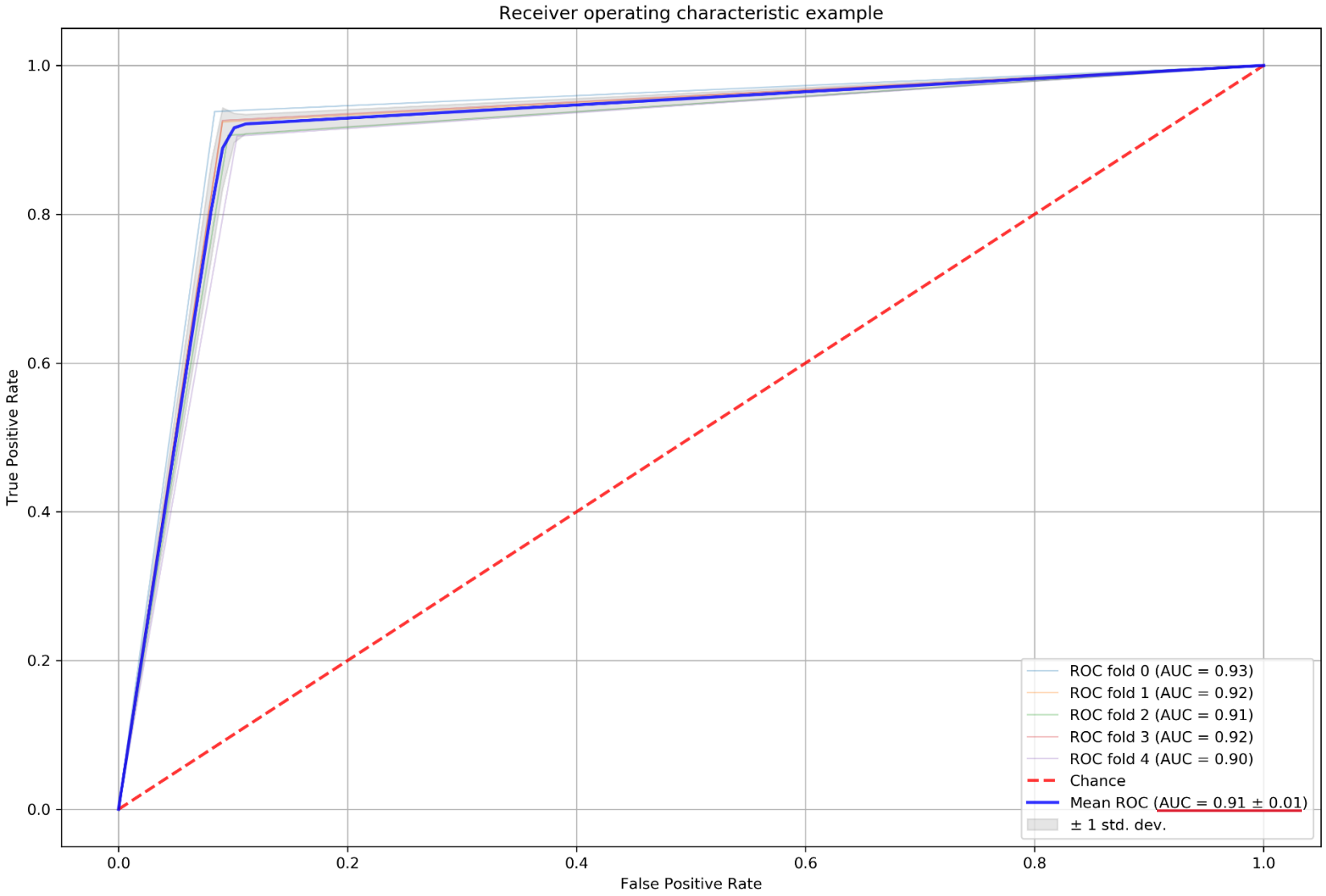


Рисунок 7. ROC кривая классификации шифрованных/сжатых данных

Результаты экспериментов III

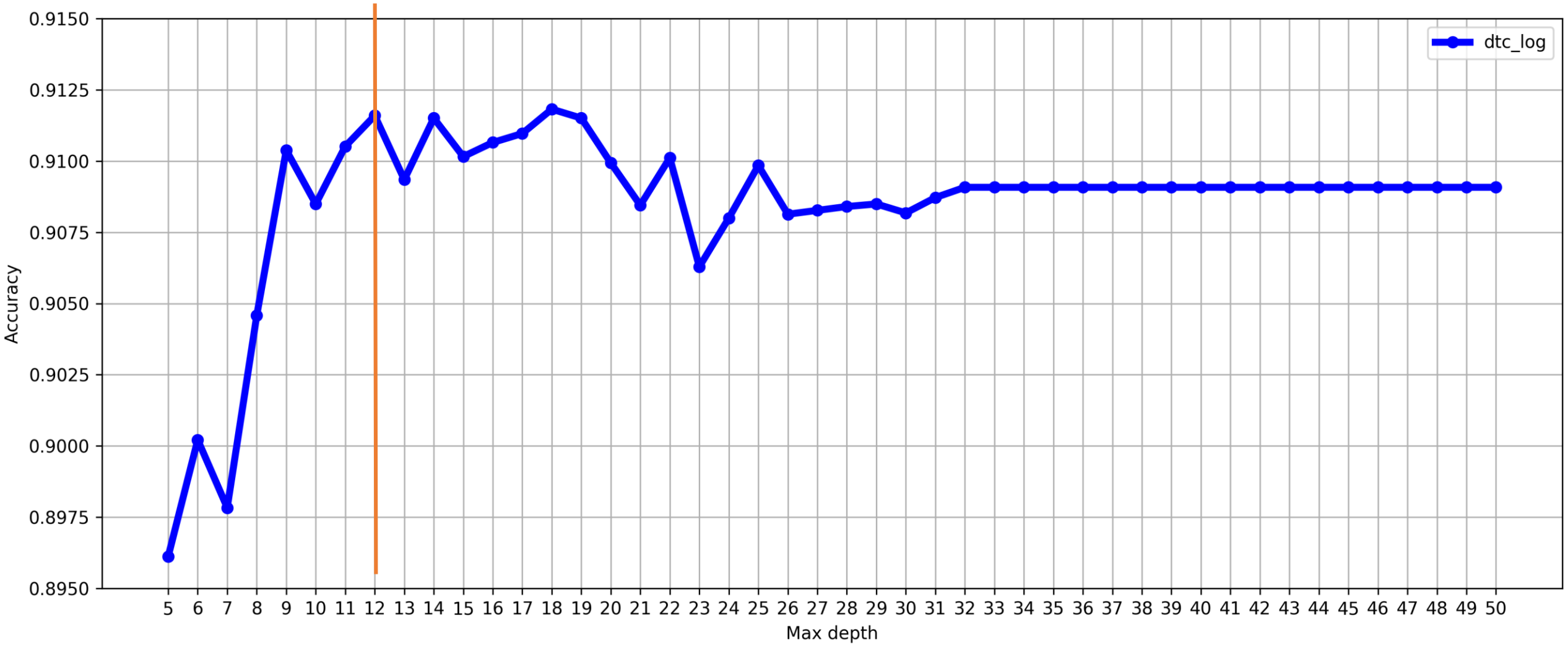


Рисунок 8. Оценка зависимости точности (Accuracy) классификации от максимальной глубины дерева решений

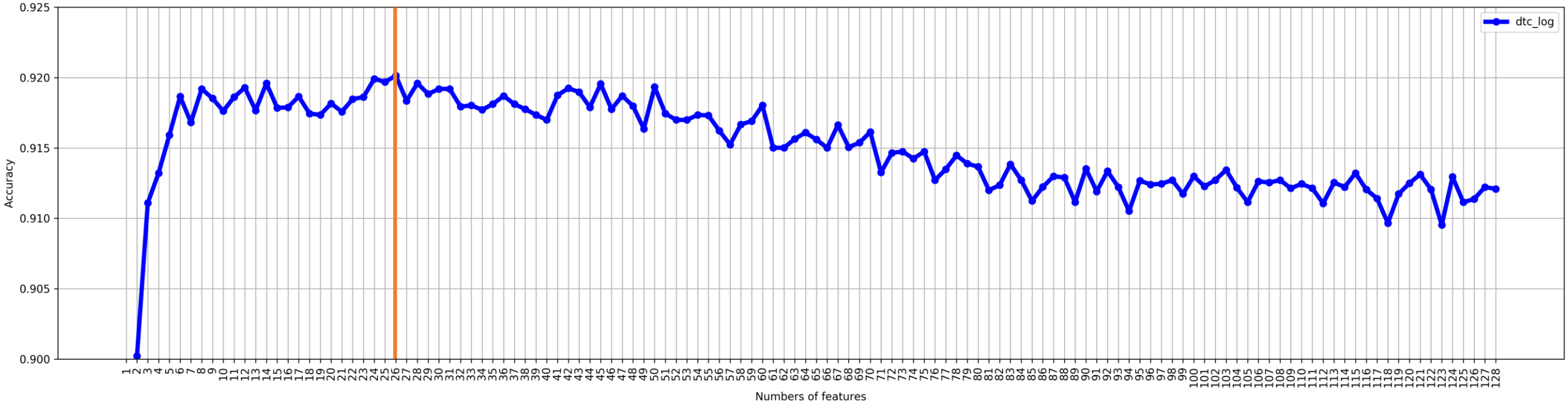
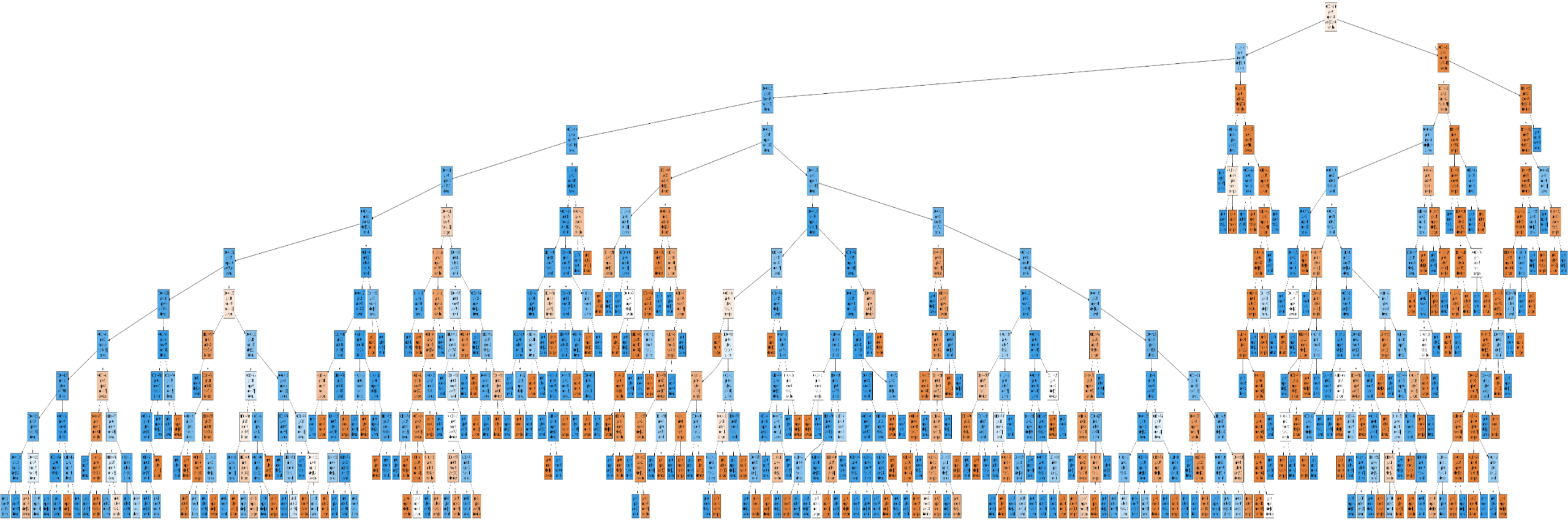


Рисунок 9. Оценка зависимости точности (Accuracy) классификации от числа признаков

Результаты экспериментов V

Гиперпараметр классификатора	Значение
Длина подпоследовательности	8 бит
Максимальная глубина дерева	12
Количество признаков	26

Мера	Значение
Accuracy	0,919
Precision	0,911
Recall	0,926



Контактная информация

Электронная почта: a.kozachok@academ.msk.rsnet.ru
Телефон: +7 920 284-57-43
Сайт: www.academ.msk.rsnet.ru

Подход к классификации последовательностей, сформированных алгоритмами сжатия и шифрования

Козачок Александр Васильевич,
д.т.н., сотрудник Академии ФСО России, г. Орёл
Спирин Андрей Андреевич,
сотрудник Академии ФСО России, г. Орёл

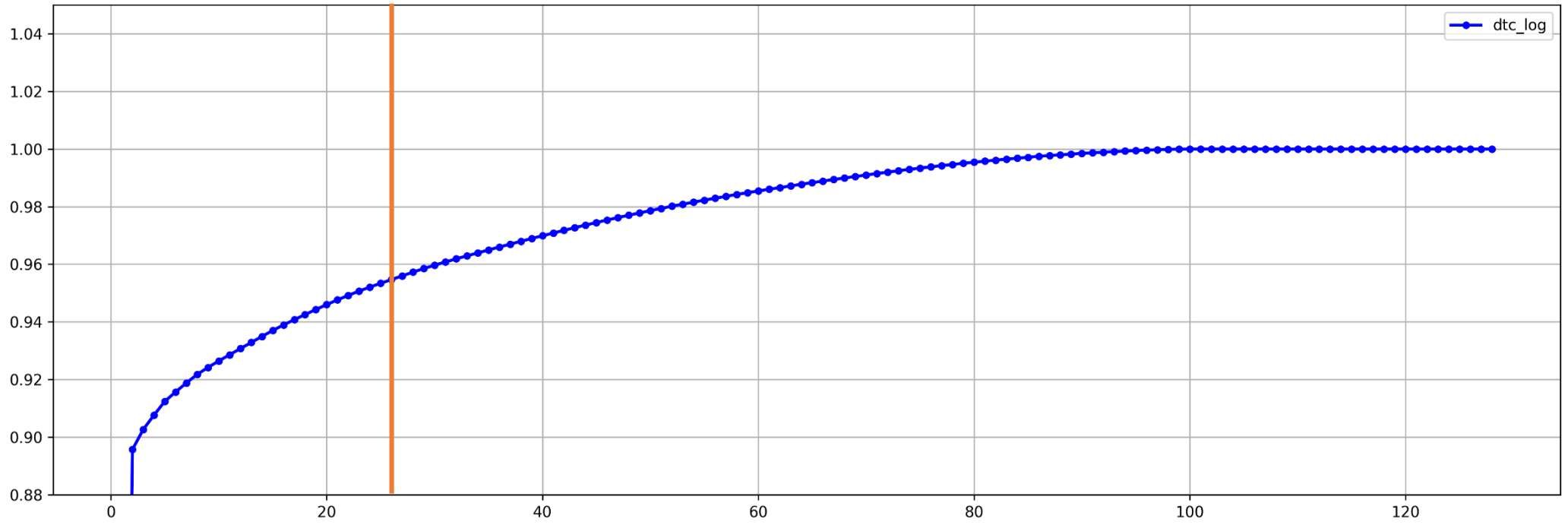


Рисунок 10. Оценка зависимости кумулятивной значимости признаков от их числа