Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«Московский государственный технический университет
имени Н. Э. Баумана
(национальный исследовательский университет)»
(МГТУ им. Н. Э. Баумана)

ФАКУЛЬТЕТ «Информатика и системы управления»

КАФЕДРА «Программное обеспечение ЭВМ и информационные технологии»

РАСЧЕТНО-ПОЯСНИТЕЛЬНАЯ ЗАПИСКА

К НАУЧНО-ИССЛЕДОВАТЕЛЬСКОЙ РАБОТЕ

НА ТЕМУ:

«Классификация методов подсчета информационной энтропии»

Студент ИУ7-73Б (Группа) _____ (Подпись, дата) Р. Р. Хамзина (И. О. Фамилия)

Руководитель НИР _____ (Подпись, дата) А. А. Оленев (И. О. Фамилия)

2022 г.

РЕФЕРАТ

Расчетно-пояснительная записка 19 с., 5 рис., 1 табл., 20 источн., 1 прил.

Объектом исследования является подсчет информационной энтропии.

Цель работы заключается в классификации методов подсчета информационной энтропии.

В рамках анализа предметной области были рассмотрены основные понятия теории информации и сжатия данных, было дано определение информационной энтропии и были представлены ее свойства.

При проведении обзора существующих методов подсчета информационной энтропии были описаны метод скользящего окна и биномиальный метод. Для их сравнения были сформулированы следующие критерии: временная сложность, необходимость вычисления факториала, возможность распараллеливания вычислений и объем требуемой дополнительной памяти.

Сравнение описанных методов по сформулированным критериям показало, что в задаче оценивания коэффициента сжатия с помощью информационной энтропии предпочтительнее использовать биномиальный метод ее подсчета.

Ключевые слова: информация, теория информации, информационная энтропия, сжатие данных, метод скользящего окна, биномиальный метод.

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

Основным ресурсом в современном обществе является информация [1]. В различных предметных областях появляются задачи, связанные с ее обработкой. Для их решения используются характеристики и подходы теории информации. Одной из таких характеристик является информационная энтропия. В лингвистике вычисление информационной энтропии применяется для определения показателей усилий, необходимых для перевода текста [2], в информационной безопасности — для оценки защищенности информационных систем [3], в медицине — для диагностики шизофрении [4] и оценки уровня анестезии [5].

С увеличением объема информации возрастают требуемый для ее хранения размер памяти и продолжительность передачи сведений. Для уменьшения размера данных и увеличения скорости их передачи используется сжатие данных [6]. В целях его оптимизации необходимо оценивать коэффициент сжатия, что может быть реализовано с помощью информационной энтропии.

Целью данной работы является классификация методов подсчета информационной энтропии.

Для достижения поставленной цели необходимо выполнить следующие задачи:

— провести анализ предметной области: рассмотреть основные определения, изучить свойства информационной энтропии и ее связь со сжатием данных;

— описать существующие методы подсчета информационной энтропии;

— выделить критерии сравнения описанных методов;

— провести сравнение методов по выделенным критериям.

1 Анализ предметной области

1.1 Основные определения

Под информацией понимают сведения, которые являются объектом хранения, передачи и обработки [7]. Формой представления информации является сообщение. Физическую величину, отображающую сообщение, называют сигналом. Передача информации осуществляется следующим образом [8]:

1. Источник информации создает случайное сообщение. В теории информации любой источник информации является стохастическим, его можно описать измеряемыми вероятностными категориям.

2. Сообщение поступает в систему передачи, в которой выполняется кодирование — преобразование сообщения с целью согласования источника информации с каналом связи для увеличения скорости передачи информации или обеспечения заданной помехоустойчивости [7]. Кодирование состоит из шифрования, сжатия и защиты от шума, в результате которых формируется сигнал.

3. Сигнал проходит через канал — среду передачи информации [8]. В канале могут возникать помехи, создаваемые источником шума.

4. Сигнал подается на вход системе приема, которая выполняет декодирование — восстановление исходного сообщения [7].

5. Исходное сообщение передается получателю.

Описанная схема передачи информации показана на рисунке 1.1.

Сообщение содержит сведения о некоторой физической системе X, которая случайным образом может перейти в какое-либо состояние x_i из конечного множества состояний $x_1, x_2, \ldots, x_n$ с вероятностями $p_1, p_2, \ldots, p_n$, где $n \in \mathbb{N}$, $p_i = P(X \sim x_i)$ и $\sum_{i=1}^{n} p_i = 1$. То есть, для такой системы существует степень неопределенности, которая описывается числом ее возможных состояний и их вероятностями. Сведения из принятого сообщения тем ценнее, чем больше была неопределенность системы до получения сообщения. Специальную

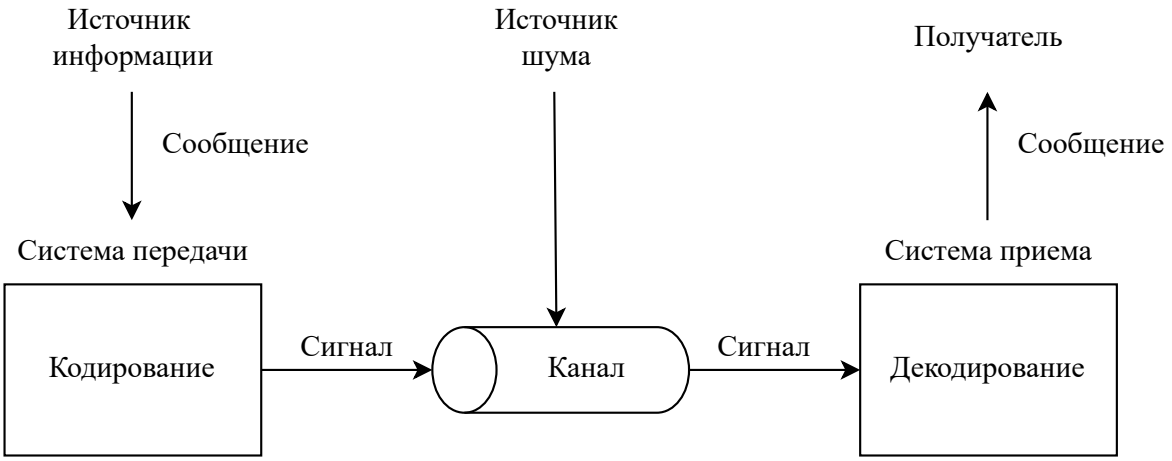

Рисунок 1.1 – Схема передачи информации

характеристику, используемую в качестве меры неопределенности системы, называют информационной энтропией [9]. Информационная энтропия конечной вероятностной схемы определяется по формуле Шеннона:

$$H(X) = -\sum_{i=1}^{n}(p_i \cdot \log_a p_i), \tag{1.1}$$

где $p_i \in [0, 1]$, $a > 1$.

Основание логарифма определяет единицы измерения информационной энтропии: при $a = 2$ энтропия измеряется в битах, при $a = 3$ — в тритах, при $a = e$ — в натах.

1.2 Свойства информационной энтропии

Информационная энтропия обладает следующими свойствами [10]:

1. Энтропия всегда неотрицательна. Значения $\log_a p_i$ в формуле (1.1) принимают неположительные значения, так как $p_i \in [0, 1]$. Поэтому

$$H(X) = -\sum_{i=1}^{n}(p_i \cdot \log_a p_i) \geqslant 0. \tag{1.2}$$

2. Энтропия равна нулю, если состояние системы в точности известно заранее. Если известно состояние x_k, в которое перейдет система X,

то вероятность этого состояния p_k равна единице, вероятности других состояний равны нулю. Тогда

$$p_k \cdot \log_a p_k = 1 \cdot \log_a 1 = 1 \cdot 0 = 0. \tag{1.3}$$

В связи с тем, что $\lim\limits_{p \to 0}(p \cdot \log_a p) = 0$, другие слагаемые суммы в формуле (1.1) равны нулю. В этом случае

$$H(X) = -\sum_{i=1}^{n}(p_i \cdot \log_a p_i) = 0. \tag{1.4}$$

3. Энтропия принимает наибольшее значение при условии, что все состояния равновероятны, то есть, $p_1 = p_2 = \cdots = p_n = \frac{1}{n}$. Тогда

$$H(X) = -\sum_{i=1}^{n}(p_i \cdot \log_a p_i) = -\sum_{i=1}^{n}(\frac{1}{n} \cdot \log_a \frac{1}{n}) = -\log_a \frac{1}{n} = \log_a n. \tag{1.5}$$

1.3 Сжатие данных

Неслучайные данные имеют некоторую структуру. Наличие у данных структуры, которую можно использовать для уменьшения их размера путем достижения такого представления данных, в котором никакая структура не выделяется, называют избыточностью. Сжатие данных — это процесс преобразования исходных данных в их компактную форму путем распознавания и использования избыточности данных [11]. Процесс сжатия состоит из двух этапов:

1. Этап моделирования, который включает в себя распознавание избыточности для построения модели. Модель представляет собой набор данных и правил, используемых для обработки входных символов.

2. Этап кодирования данных с использованием модели.

Описанные этапы сжатия данных представлены на рисунке 1.2.

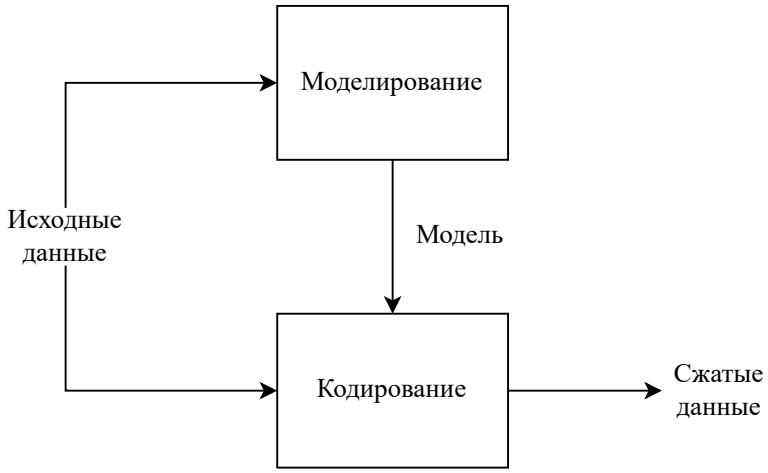

Рисунок 1.2 – Этапы сжатия данных

В результате сжатия исходных данных X получается их представление $X_{\text{сж}}$. При восстановлении сжатые данные $X_{\text{сж}}$ преобразуются в представление Y. На основании требований к восстановлению выделяют [12]:

— сжатие данных без потерь, при котором $Y = X$;

— сжатие данных с потерями, при котором $Y \neq X$.

Схемы сжатия данных без потерь и с потерями показаны на рисунке 1.3.

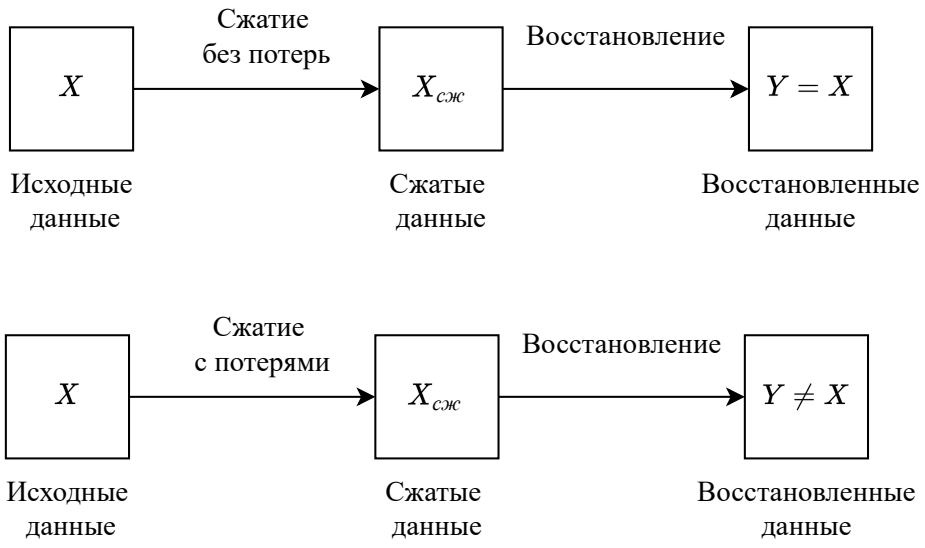

Рисунок 1.3 – Схемы сжатия данных без потерь и с потерями

Для измерения производительности сжатия используют характеристику, которую называют коэффициентом сжатия и определяют следующим образом [13]:

$$K_{\text{сж}} = \frac{L_{\text{исх}}}{L_{\text{сж}}}, \tag{1.6}$$

где $L_{\text{исх}}$ — объем исходных данных, $L_{\text{сж}}$ — объем сжатых данных.

Данные, обладающие предсказуемой структурой, сокращают неопределенность системы меньше, чем сведения, в которых никакая структура не выделяется. Так как информационная энтропия является мерой неопределенности системы, то данные с выделяемой структурой, имеют низкое значение энтропии. Сведения, в которых закономерности не определяются, имеют высокое значение энтропии [14]. Так, чем меньше избыточность данных, тем выше значение их энтропии. То есть, информационная энтропия сжатых данных выше, чем ее значение до сжатия.

Согласно теореме Шеннона об источнике шифрования сигнал, обладающий размером S и информационной энтропией H, не может быть сжат менее, чем до $S \cdot H$ битов без потери точности информации. Таким образом, на основании информационной энтропии исходных данных определяется теоретическая граница коэффициента сжатия [15].

В связи с применением операций сложения, умножения и логарифмирования при вычислении энтропии по формуле (1.1), число которых растет с увеличением объема данных, встает задача выбора метода подсчета. Использование метода влияет на скорость и время определения информационной энтропии.

2 Описание существующих методов подсчета

2.1 Метод скользящего окна

При решении задач, связанных со сжатием, данные X представляют собой массив байтов размером N [16]. Байт состоит из восьми битов, каждый из которых кодирует одно из значений множества $\{0, 1\}$. Поэтому один байт может принимать значения из интервала от 0 до 255 включительно в десятичной системе счисления.

В методе скользящего окна под окном понимают рассматриваемую на текущем этапе подпоследовательность данных размером n [17]. При подсчете энтропии данным методом необходима дополнительная память размером 2^n для хранения числа вхождений подпоследовательностей данных. Так как с увеличением размера окна, растет объем дополнительной памяти, в качестве окна выбирается минимально адресуемая единица памяти, которой является байт [18]. В связи с тем, что на каждом этапе окно смещается на следующие восемь битов, как показано на рисунке 2.1, оно называется скользящим.

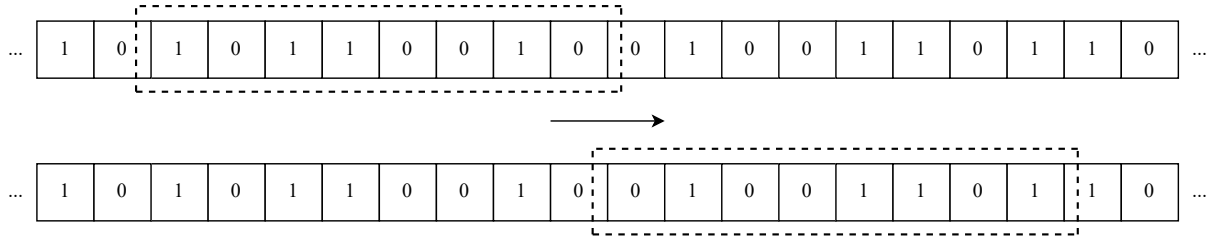

Рисунок 2.1 – Проход по массиву байтов методом скользящего окна

Вычисление информационной энтропии методом скользящего окна включает в себя два шага:

1. Для каждого возможного значения байта подсчитывается число его вхождений k_i в массив байтов, где $i = \overline{0, 255}$.

2. С учетом того, что вероятность появления байта в массиве $p_i = \frac{k_i}{N}$, информационная энтропия вычисляется по следующей формуле:

$$H(X) = -\sum_{i=0}^{255}(p_i \cdot \log_2 p_i). \tag{2.1}$$

Так как первый шаг метода предполагает проход по массиву байтов размером N, а второй шаг — проход по массиву числа вхождений подпоследовательностей данных размером 2^n, временная сложность метода скользящего окна — $O(N + 2^n)$.

Информационная энтропия, подсчитанная методом скользящего окна, принимает значения из интервала $[0, 8]$ битов. При этом согласно свойству 2 из раздела 1.2 информационная энтропия принимает нулевое значение в случае, когда массив данных состоит из одинаковых байтов, и в соответствии со свойством 3 из раздела 1.2 максимальное значение, равное восьми, если все байты в массиве различны.

2.2 Биномиальный метод

В биномиальном методе вычисления информационной энтропии [19] данные X рассматриваются как последовательность сообщений, генерируемых бернуллиевским источником — источником информации, порождающим символы из алфавита $\{0; 1\}$ с вероятностями $1 - p$ и p соответственно, причем $p \in (0, 1)$ и может быть неизвестно [20]. То есть, сообщение представляет собой последовательность битов длины n.

Вероятность того, что сообщение содержит k единиц, где $k = \overline{0, n}$, вычисляется следующим образом:

$$P_k = p^k \cdot (1 - p)^{(n-k)}. \tag{2.2}$$

Количество возможных сообщений, содержащих k единиц, определяется как биномиальный коэффициент:

$$C_n^k = \frac{n!}{k! \cdot (n - k)!}. \tag{2.3}$$

В связи с тем, что k может принимать значения из интервала $[0, n]$, число биномиальных коэффициентов, подсчитанных по формуле (2.3), равно $n + 1$. Это означает, что сообщения можно разбить на $n+1$ классов эквивалентности. Тогда информационная энтропия вычисляется так:

$$H(X) = -\sum_{k=0}^{n}(C_n^k \cdot P_k \cdot \log_2 P_k). \tag{2.4}$$

В соответствии с формулой (2.4) биномиальный метод подсчета информационной энтропии состоит из следующих этапов:

1. Исходные сообщения разбиваются на $n + 1$ классов эквивалентности, сообщения которых содержат $k = \overline{0, n}$ единиц.

2. Для каждого класса эквивалентности рассчитывается биномиальный коэффициент C_n^k.

3. Определяются вероятности p появления единицы в сообщениях.

4. Вычисляются вероятности P_k по формуле (2.2).

5. Суммируются произведения биномиальных коэффициентов C_n^k, вероятностей P_k и логарифмов вероятностей P_k для всех $n + 1$ значений k.

Схема определения биномиальных коэффициентов C_n^k и вероятностей P_k представлена на рисунке 2.2.

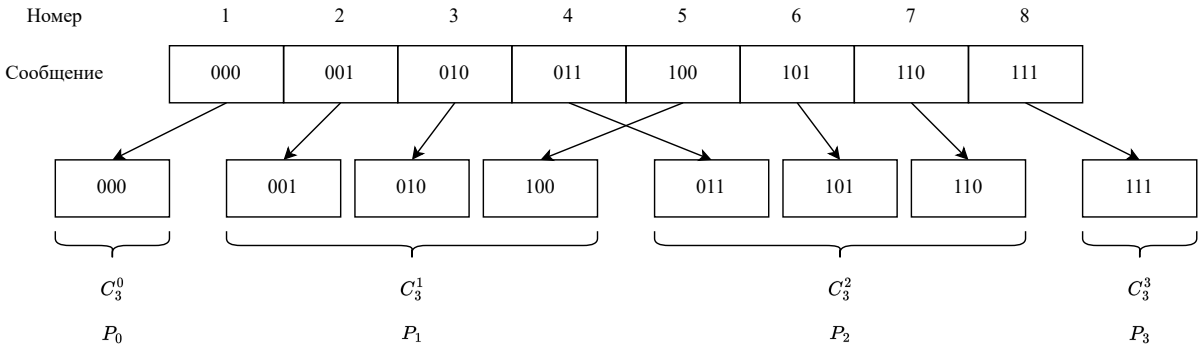

Рисунок 2.2 – Определение биномиальных коэффициентов C_n^k и вероятностей P_k

Для определения вероятности p появления единицы в сообщениях на третьем этапе биномиального метода необходима дополнительная память размером $n+1$. При этом данный этап подразумевает проход по массиву байтов размером N. Последний этап вычисления предполагает проход по массиву, содержащему значения вероятностей p появления единицы в сообщениях. Тогда временная сложность биномиального метода подсчета информационной энтропии — $O(N + n)$.

В биномиальном методе определяются вероятности появления не подпоследовательности, а единицы в подпоследовательностях. Кроме того, расчет ускоряется за счет разделения исходной последовательности на классы эквивалентности, что приводит к сокращению операций сложения.

При рассмотрении бернуллиевского источника информации предполагается, что вероятность появления единицы в подпоследовательности не зависит от вероятностей появления нуля или единицы в битах предыдущей подпоследовательности. При наличии такой зависимости вычисленное значение энтропии будет завышено. Так как рассматриваемое представление данных — последовательность битов, то вероятности появления каждого значения бита независимы.

Недостатком данного метода является трудоемкость вычисления факториала при определении биномиальных коэффициентов с увеличением длины подпоследовательности. Для снижения времени подсчета биномиальных коэффициентов можно хранить их значения в дополнительном массиве. Так как для сообщения длины n количество требуемых для определения энтропии биномиальных коэффициентов равно $n + 1$, то для их хранения потребуется память размером $n + 1$.

3 Классификация существующих методов подсчета

При вычислении информационной энтропии методом скользящего окна и биномиальным методом операции сложения, умножения и логарифмирования применяются к целым числам и числам с плавающей запятой.

Для сравнения методов подсчета информационной энтропии были выделены следующие критерии оценки:

— К1 — временная сложность;

— К2 — необходимость вычисления факториала;

— К3 — возможность распараллеливания вычислений;

— К4 — объем требуемой дополнительной памяти.

Результаты сравнения представлены в таблице 3.1.

Таблица 3.1 – Сравнение методов подсчета информационной энтропии

Метод	**К1**	**К2**	**К3**	**К4**
Скользящего окна	$O(N + 2^n)$	$-$	$+$	2^n
Биномиальный	$O(N + n)$	$+$	$+$	$2 \cdot (n + 1)$

Таким образом, биномиальный метод подсчета информационной энтропии требует меньших вычислительных затрат по времени и по памяти.

ЗАКЛЮЧЕНИЕ

В ходе выполнения научно-исследовательской работы были классифицированы методы подсчета информационной энтропии.

На основании результатов сравнения методов можно сделать вывод о том, что в задаче оценивания коэффициента сжатия с помощью информационной энтропии предпочтительнее использовать биномиальный метод ее подсчета.

При написании данной работы:

— проведен анализ предметной области: рассмотрены основные определения, изучены свойства информационной энтропии и ее связь со сжатием данных;

— описаны существующие методы подсчета информационной энтропии;

— выделены критерии сравнения описанных методов;

— проведено сравнение методов по выделенным критериям.

Таким образом, поставленные задачи были выполнены, цель научно-исследовательской работы была достигнута.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. *Коржова Д. А.*, *Смеричевский Э. Ф.* Информационное общество: к анализу понятия // Вестник Науки и Творчества. — 2019. — № 7(43). — 9—13 с.
2. *Carl M.*, *Tonge A.*, *Lacruz I.* A systems theory perspective on the translation process // Translation, Cognition & Behavior. — 2019. — 211—232 с.
3. *Imanbayeva A.* Evaluating the effectiveness of information security based on the calculation of information entropy // Journal of Physics: Conference Series. — 2021. — 12—42 с.
4. *Кутепов И. Е.* Визуализация энтропии сигналов ЭЭГ при шизофрении // Научная визуализация. — 2020. — 1—9 с.
5. *Mohammad O.*, *Amanbaeva G. M.* Entropy monitoring in medicine // Eurasian Medical Journal. — 2020. — № 2. — 28—32 с.
6. *Anup A.*, *Ashok R.*, *Raundale P.* Comparative Study of Data Compression Techniques // International Journal of Computer Applications. — 2019. — 15—19 с.
7. *Березкин Е. Ф.* Основы теории информации и кодирования: учебное пособие // 3-е изд., стер. — СПб.: Лань. — 2022. — 320 с.
8. *Rodrigues M.* Information-Theoretic Methods in Data Science // Cambridge: Cambridge University Press. — 2021. — 43 с.
9. *Попов И. Ю.*, *Блинова И. В.* Теория информации // 3-е изд., стер. — СПб.: Лань. — 2022. — 160 с.
10. *Осокин А. Н.*, *Мальчуков А. Н.* Теория информации: учебное пособие для вузов // М.: Издательство Юрайт. — 2022. — 205 с.
11. *Uthayakumar J.*, *Vengattaraman T.*, *Dhavachelvan P.* A survey on data compression techniques: From the perspective of data quality, coding schemes, data type and applications // Journal of King Saud University — Computer and Information Sciences. — 2021. — 119—140 с.
12. *Пантелеев Е. Р.*, *Алыкова А. Л.* Алгоритмы сжатия данных без потерь: учебное пособие для вузов // 2-е изд., стер. — СПб.: Лань. — 2022. — 172 с.

13. *Gupta A.*, *Nigam S.* A Review on Different Types of Lossless Data Compression Techniques // International Journal of Scientific Research in Computer Science, Engineering and Information Technology. — 2021. — 50—56 с.

14. *Zbili M.*, *Rama S.* A Quick and Easy Way to Estimate Entropy and Mutual Information for Neuroscience // Frontiers in Neuroinformatics. — 2021.

15. *Cheng X.*, *Li Z.* How does Shannon's source coding theorem fare in prediction of image comrpession ratio with current algorithms? // International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences. — 2020. — 1313—1319 с.

16. *Ryabko B.* Time-Universal Data Compression // Algorithms. — 2019.

17. *Guo H.* File Entropy Signal Analysis Combined With Wavelet Decomposition for Malware Classification // IEEE Access. — 2020. — 158961—158971 с.

18. *Пухальский Г. И.*, *Новосельцева Т. Я.* Проектирование цифровых устройств: учебное пособие для вузов // СПб.: Лань. — 2022. — 896 с.

19. *Borysenko O.* On the binomial method for calculation of entropy // Grail of Science. — 2022. — 113—118 с.

20. *Рябко Б. Я.*, *Фионов А. Н.* Криптография в информационном мире // М.: Горячая линия-Телеком. — 2018. — 300 с.

ПРИЛОЖЕНИЕ А