Лабораторная работа №5

Кодирование по Шеннону–Фано (статистический префиксный код)

Лабораторная работа №5

Дисциплина: Основы кодирования

Тема: Оптимальное кодирование информации. Код Шеннона–Фано

1. Цель работы

Целью лабораторной работы является изучение принципов оптимального кодирования информации, детальное освоение метода Шеннона–Фано и анализ эффективности данного метода по сравнению с равномерным и простейшими неравномерными кодами.

2. Исходные данные

В лабораторной работе используются данные, полученные в предыдущих лабораторных работах. Формирование новых исходных данных не допускается.

Студент использует:
- источник информации, сформированный из ФИО студента и даты рождения (при необходимости — с добавлением данных членов семьи);
- алфавит источника информации;
- вероятности появления символов p_i;
- энтропию источника информации H.

3. Теоретические сведения

В теории информации под оптимальным кодированием понимается такое кодирование, при котором средняя длина кодового слова минимальна или близка к минимально возможной.

Префиксный код — код, в котором ни одно кодовое слово не является началом (префиксом) другого. Префиксность гарантирует однозначное декодирование без разделителей.

Средняя длина кода:

$$\bar{L} = \Sigma\, p_i \cdot l_i,$$

где p_i — вероятность символа, l_i — длина его кодового слова.

Энтропия источника:

$$H = -\Sigma\, p_i \cdot \log_2(p_i).$$

Выполняется нижняя граница: $H \le \bar{L}$ для любого однозначно декодируемого кода.

Эффективность:

$$\eta = H/\bar{L},$$

избыточность:

$$r = 1 - \eta.$$

Чем ближе $\bar{L}$ к H, тем лучше код.

Код Шеннона–Фано является префиксным кодом и строится с учётом вероятностей символов: более вероятные символы получают более короткие коды. Он обеспечивает среднюю длину кода, близкую к энтропии источника, однако не всегда является строго оптимальным. Код Шеннона–Фано относится к статистическим методам кодирования. Он использует вероятности символов источника и строит префиксный двоичный код, в котором более вероятным символам назначаются более короткие кодовые слова. В отличие от Хаффмана, Шеннон–Фано не всегда даёт оптимум, но часто даёт заметное уменьшение $\bar{L}$ по сравнению с равномерным кодом.

Алгоритм Шеннона–Фано

Алгоритм построения кода Шеннона–Фано включает следующие шаги:

1. Символы алфавита упорядочиваются по убыванию вероятностей.
2. Упорядоченное множество символов разбивается на две группы так, чтобы суммы вероятностей групп были максимально близки.
3. Первой группе присваивается бит 0, второй — бит 1.
4. Процедура рекурсивно повторяется для каждой группы до тех пор, пока каждой группе не будет соответствовать один символ.

Пример построения кода Шеннона–Фано

Рассмотрим источник информации с алфавитом:
A (0.40), B (0.30), C (0.20), D (0.10).

1. Символы упорядочены по убыванию вероятностей.
2. Множество делится на группы {A, B} с суммарной вероятностью 0.70 и {C, D} с суммарной вероятностью 0.30.
3. Группе {A, B} присваивается 0, группе {C, D} — 1.
4. Далее выполняется разбиение внутри каждой группы, в результате чего формируется префиксный код.

Полученный код будет короче для более вероятных символов.

4. Задания

Все задания выполняются последовательно: A → B → C.

Уровень A. Подготовка данных

A.1. Отсортировать символы алфавита по убыванию вероятностей p_i.
A.2. Сформировать таблицу вероятностей и проверить выполнение условия
$\Sigma p_i = 1$.

Уровень B. Построение кода Шеннона–Фано

B.1. Выполнить разбиение множества символов на группы с близкими суммарными вероятностями.
B.2. Построить код Шеннона–Фано с использованием рекурсивного алгоритма.
B.3. Сформировать итоговую таблицу: символ – вероятность – код – длина кода.

Уровень C. Анализ эффективности кодирования

C.1. Рассчитать среднюю длину кода Шеннона–Фано.
C.2. Сравнить среднюю длину кода с энтропией источника и со средней длиной равномерного кода.
C.3. Сделать выводы об эффективности и ограничениях метода Шеннона–Фано.

5. Пример

Исходные данные (как в ЛР1–ЛР4)

Студент: АЛПАМЫССЕРИКНУРАЛЫЕВИЧ12042000

Мать: ДАНИЯАЙГУЛСАПАРОВНА05071975

Отец: ЕРНАРТАЛГАТМУХТАРОВИЧ23121970

Объединённое сообщение (источник):

АЛПАМЫССЕРИКНУРАЛЫЕВИЧ12042000ДАНИЯАЙГУЛСАПАРОВНА05071975ЕРНАРТАЛГАТМУХТАРОВИЧ23121970

Длина сообщения: N = 86 символов.

Мощность алфавита источника по факту: |A| = 29 символов (кириллица + цифры).

5.1 Задание 1 — решение: таблица частот и вероятностей (из ЛР3)

Составляем алфавит и считаем частоты f_i, вероятности $p_i = f_i/N$. Порядок оставляем таким же, как в предыдущих работах, чтобы сохранить преемственность.

Символ	f_i	$p_i = f_i/N$	$-\log_2(p_i)$
А	12	0.139535	2.841302
0	7	0.081395	3.618910
Р	6	0.069767	3.841302
1	4	0.046512	4.426265
2	4	0.046512	4.426265
И	4	0.046512	4.426265
Л	4	0.046512	4.426265
Н	4	0.046512	4.426265
7	3	0.034884	4.841302
В	3	0.034884	4.841302
Е	3	0.034884	4.841302
С	3	0.034884	4.841302
Т	3	0.034884	4.841302
У	3	0.034884	4.841302
5	2	0.023256	5.426265
9	2	0.023256	5.426265
Г	2	0.023256	5.426265
М	2	0.023256	5.426265
О	2	0.023256	5.426265
П	2	0.023256	5.426265
Ч	2	0.023256	5.426265
Ы	2	0.023256	5.426265
З	1	0.011628	6.426265
4	1	0.011628	6.426265
Д	1	0.011628	6.426265
Й	1	0.011628	6.426265
К	1	0.011628	6.426265
Х	1	0.011628	6.426265
Я	1	0.011628	6.426265

Проверка: $\Sigma f_i = 86 = N$, $\Sigma p_i = 1.000000$.

5.2. Задание 2 — решение: построение кода Шеннона–Фано

Алгоритм:

Шаг 1) Упорядочиваем символы по убыванию вероятностей.

Шаг 2) Делим список на две группы с максимально близкими суммарными вероятностями.

Шаг 3) Левой группе приписываем бит 0, правой — бит 1.

Шаг 4) Повторяем разбиение рекурсивно в каждой группе, пока не останется по одному символу.

Ниже приведены первые ключевые разбиения (идея «как на доске»). Полный список разбиений — в Приложении А.

Префикс	Группа (символы)	Σp	0-группа	Σp(0)	1-группа	Σp(1)

∅	A, 0, P, 1, 2, И, Л, Н, |, 7, В, Е, С, Т, У, 5, 9, Г, М, О, П, Ч, Ы, 3, 4, Д, Й, К, Х, Я	1.000000	A, 0, P, 1, 2, И, Л, Н	0.523256	7, В, Е, С, Т, У, 5, 9, Г, М, О, П, Ч, Ы, 3, 4, Д, Й, К, Х, Я	0.476744
0	A, 0, P, |, 1, 2, И, Л, Н	0.523256	A, 0, P	0.290698	1, 2, И, Л, Н	0.232558
00	A, |, 0, P	0.290698	A	0.139535	0, P	0.151163
001	0, |, P	0.151163	0	0.081395	P	0.069767
01	1, 2, |, И, Л, Н	0.232558	1, 2	0.093023	И, Л, Н	0.139535
010	1, |, 2	0.093023	1	0.046512	2	0.046512
011	И, |, Л, Н	0.139535	И	0.046512	Л, Н	0.093023
0111	Л, |, Н	0.093023	Л	0.046512	Н	0.046512
1	7, В, Е, С, Т, У, 5, |, 9, Г, М, О, П, Ч, Ы, 3, 4, Д, Й, К, Х, Я	0.476744	7, В, Е, С, Т, У, 5	0.232558	9, Г, М, О, П, Ч, Ы, 3, 4, Д, Й, К, Х, Я	0.244186
10	7, В, Е, |, С, Т, У, 5	0.232558	7, В, Е	0.104651	С, Т, У, 5	0.127907

∅ = пустой префикс (начальный уровень дерева)

После завершения рекурсивных разбиений каждому символу соответствует уникальный двоичный код. Код является префиксным.

Эталонная таблица кодов Шеннона–Фано с побитовой структурой.

Символ	f_i	p_i	1 бит	2 бит	3 бит	4 бит	5 бит	6 бит	7 бит	l_i
A	12	0.139535	0	0	0					3
0	7	0.081395			1	0				4
P	6	0.069767				1				4
1	4	0.046512		1	0	0				4
2	4	0.046512				1				4
И	4	0.046512			1	0				4
Л	4	0.046512				1	0			5
Н	4	0.046512					1			5
7	3	0.034884	1	0	0	0				4
В	3	0.034884				1	0			5
Е	3	0.034884					1			5
С	3	0.034884			1	0	0			5
Т	3	0.034884					1			5
У	3	0.034884				1	0			5

5	2	0.023256					1			5
9	2	0.023256		1	0	0	0			5
Г	2	0.023256					1			5
М	2	0.023256				1	0			5
О	2	0.023256					1	0		6
П	2	0.023256						1		6
Ч	2	0.023256			1	0	0			5
Ы	2	0.023256					1	0		6
З	1	0.011628						1		6
4	1	0.011628				1	0	0		6
Д	1	0.011628						1	0	7
Й	1	0.011628							1	7
К	1	0.011628					1	0		6
Х	1	0.011628						1	0	7
Я	1	0.011628							1	7

5.3 Задание 3 — решение: итоговая таблица кода Шеннона–Фано

Формируем итоговую таблицу: символ, f_i, p_i, код, длина l_i.

Символ	f_i	p_i	Код Шеннона–Фано	l_i
А	12	0.139535	000	3
0	7	0.081395	0010	4
Р	6	0.069767	0011	4
1	4	0.046512	0100	4
2	4	0.046512	0101	4
И	4	0.046512	0110	4
Л	4	0.046512	01110	5
Н	4	0.046512	01111	5
7	3	0.034884	1000	4
В	3	0.034884	10010	5
Е	3	0.034884	10011	5
С	3	0.034884	10100	5
Т	3	0.034884	10101	5
У	3	0.034884	10110	5
5	2	0.023256	10111	5
9	2	0.023256	11000	5
Г	2	0.023256	11001	5
М	2	0.023256	11010	5
О	2	0.023256	110110	6
П	2	0.023256	110111	6
Ч	2	0.023256	11100	5
Ы	2	0.023256	111010	6
З	1	0.011628	111011	6
4	1	0.011628	111100	6
Д	1	0.011628	1111010	7
Й	1	0.011628	1111011	7

К	1	0.011628	111110	6
Х	1	0.011628	1111110	7
Я	1	0.011628	1111111	7

5.4. Задание 4 — решение: кодирование сообщения и длина (B)

Кодируем исходное сообщение, заменяя каждый символ его кодовым словом из таблицы.

Чтобы продемонстрировать принцип, приведём кодирование первых 20 символов:

Первые 20 символов: АЛПАМЫССЕРИКНУРАЛЫЕВ

Коды по порядку: 000 01110 110111 000 11010 111010 10100 10100 10011 0011 0110 111110 01111 10110 0011 000 01110 111010 10011 10010

Первые биты (слитно): 00001110110111000110101110101010010100100110011011011111001111101100011000011101110101001110010

Длина фрагмента (20 символов): 95 бит.

Полный битовый поток в отчёт обычно не вставляют целиком из-за объёма; вместо этого обязательно приводят итоговую длину.

Итоговая длина закодированного сообщения по Шеннону–Фано:

$B\_SF = \Sigma f_i \cdot l_i = 395$ бит.

5.5. Задание 5 — решение: средняя длина кода ($\bar{L}$)

Средняя длина кода вычисляется двумя эквивалентными способами:

1) $\bar{L} = \Sigma p_i \cdot l_i$

2) $\bar{L} = B/N$

По таблице: $\bar{L}\_SF = \Sigma p_i \cdot l_i = 4.593023$ бит/символ.

Проверка через B/N: $\bar{L}\_SF = 395/86 = 4.593023$ бит/символ.

5.6 Задание 6 — решение: сравнение с равномерным кодом (ЛР4)

В ЛР4 равномерный двоичный код имел фиксированную длину L = 5 бит (так как $\lceil \log_2 29 \rceil = 5$).

Равномерный код: $B\_uniform = N \cdot 5 = 430$ бит, $\bar{L}\_uniform = 5.000000$ бит/символ.

Шеннон–Фано: $B\_SF = 395$ бит, $\bar{L}\_SF = 4.593023$ бит/символ.

Выигрыш по длине сообщения: $\Delta B = 430 - 395 = 35$ бит.

Относительный выигрыш: 8.14% (по сравнению с равномерным кодом).

5.7 Задание 7 — решение: сравнение с кодом Шеннона из ЛР3

В ЛР3 использовалось правило Шеннона: $l_i = \lceil -\log_2(p_i) \rceil$ и была получена суммарная длина $B\_Shannon = 423$ бит.

Код Шеннона (ЛР3): $B\_Shannon = 423$ бит, $\bar{L}\_Shannon = 4.918605$ бит/символ.

Шеннон–Фано: $B\_SF = 395$ бит, $\bar{L}\_SF = 4.593023$ бит/символ.

Выигрыш Шеннона–Фано относительно кода Шеннона: $\Delta B = 423 - 395 = 28$ бит (6.62%).

5.8. Задание 8 — решение: энтропия, эффективность и избыточность

Энтропия источника (по вероятностям из таблицы):

$H = -\Sigma p_i \cdot \log_2(p_i) = 4.534286$ бит/символ.

Метод	B (бит)	$\bar{L}$ (бит/символ)	$\eta = H/\bar{L}$	$r = 1-\eta$
Равномерный (ЛР4)	430	5.000000	0.906857	0.093143
Шеннон (ЛР3)	423	4.918605	0.921864	0.078136

Шеннон–Фано (ЛР5)	395	4.593023	0.987212	0.012788

Чем ближе $\bar{L}$ к H, тем выше эффективность. В данном примере Шеннон–Фано существенно приближает $\bar{L}$ к H.

5.8. Задание 9 — итоговый вывод

1) Равномерное кодирование (ЛР4) не использует статистику символов и даёт $\bar{L} = 5$ бит/символ.
2) Код Шеннона–Фано использует вероятности и строит префиксный код, получая $\bar{L} \approx 4.593$ бит/символ и B = 395 бит.
3) В сравнении с равномерным кодом длина сообщения уменьшилась на 35 бит (8.14%).
4) Эффективность Шеннон–Фано $\eta \approx 0.987$, избыточность $r \approx 0.013$.
5) Полученный результат логично подводит к следующему шагу курса — оптимальному кодированию по Хаффману (ЛР6).

Приложение А. Полный список разбиений (дерево Шеннона–Фано)

Ниже приведены все разбиения групп на 0- и 1-подгруппы до единичных символов (всего n−1 разбиений).

№	Префикс	Символы группы	0-группа (Σp)	1-группа (Σp)	Δ
1	∅	А, 0, Р, 1, 2, И, Л, Н, |, 7, В, Е, С, Т, У, 5, 9, Г, М, О, П, Ч, Ы, 3, 4, Д, Й, К, Х, Я	А, 0, Р, 1, 2, И, Л, Н (0.523256)	7, В, Е, С, Т, У, 5, 9, Г, М, О, П, Ч, Ы, 3, 4, Д, Й, К, Х, Я (0.476744)	0.023256
2	0	А, 0, Р, |, 1, 2, И, Л, Н	А, 0, Р (0.290698)	1, 2, И, Л, Н (0.232558)	0.029070
3	00	А, |, 0, Р	А (0.139535)	0, Р (0.151163)	0.005814
4	001	0, |, Р	0 (0.081395)	Р (0.069767)	0.005814
5	01	1, 2, |, И, Л, Н	1, 2 (0.093023)	И, Л, Н (0.139535)	0.023256
6	010	1, |, 2	1 (0.046512)	2 (0.046512)	0.000000
7	011	И, |, Л, Н	И (0.046512)	Л, Н (0.093023)	0.023256
8	0111	Л, |, Н	Л (0.046512)	Н (0.046512)	0.000000
9	1	7, В, Е, С, Т, У, 5, |, 9, Г, М, О, П, Ч, Ы, 3, 4, Д, Й, К, Х, Я	7, В, Е, С, Т, У, 5 (0.232558)	9, Г, М, О, П, Ч, Ы, 3, 4, Д, Й, К, Х, Я (0.244186)	0.005814

10	10	7, В, Е, |, С, Т, У, 5	7, В, Е (0.104651)	С, Т, У, 5 (0.127907)	0.011628
11	100	7, |, В, Е	7 (0.034884)	В, Е (0.069767)	0.017442
12	1001	В, |, Е	В (0.034884)	Е (0.034884)	0.000000
13	101	С, Т, |, У, 5	С, Т (0.069767)	У, 5 (0.058140)	0.005814
14	1010	С, |, Т	С (0.034884)	Т (0.034884)	0.000000
15	1011	У, |, 5	У (0.034884)	5 (0.023256)	0.005814
16	11	9, Г, М, О, П, |, Ч, Ы, 3, 4, Д, Й, К, Х, Я	9, Г, М, О, П (0.116279)	Ч, Ы, 3, 4, Д, Й, К, Х, Я (0.127907)	0.005814
17	110	9, Г, |, М, О, П	9, Г (0.046512)	М, О, П (0.069767)	0.011628
18	1100	9, |, Г	9 (0.023256)	Г (0.023256)	0.000000
19	1101	М, |, О, П	М (0.023256)	О, П (0.046512)	0.011628
20	11011	О, |, П	О (0.023256)	П (0.023256)	0.000000
21	111	Ч, Ы, 3, |, 4, Д, Й, К, Х, Я	Ч, Ы, 3 (0.058140)	4, Д, Й, К, Х, Я (0.069767)	0.005814
22	1110	Ч, |, Ы, 3	Ч (0.023256)	Ы, 3 (0.034884)	0.005814
23	11101	Ы, |, 3	Ы (0.023256)	3 (0.011628)	0.005814
24	1111	4, Д, Й, |, К, Х, Я	4, Д, Й (0.034884)	К, Х, Я (0.034884)	0.000000
25	11110	4, |, Д, Й	4 (0.011628)	Д, Й (0.023256)	0.005814
26	111101	Д, |, Й	Д (0.011628)	Й (0.011628)	0.000000
27	11111	К, |, Х, Я	К (0.011628)	Х, Я (0.023256)	0.005814
28	111111	Х, |, Я	Х (0.011628)	Я (0.011628)	0.000000

6. Требования к программе на Python (продолжение проекта ЛР1–ЛР4)

Разрешается и рекомендуется дописывать программу из ЛР1–ЛР4 (единый проект). Программа должна:

1) Принимать исходные строки (ФИО+дата) для студента и членов семьи и формировать объединённое сообщение.

2) Считать частоты f_i, вероятности p_i.
3) Строить код Шеннона–Фано по алгоритму разбиений.
4) Выводить таблицу: символ, f_i, p_i, код, l_i.
5) Кодировать сообщение и выводить B, $\bar{L}$.
6) Считать H, η, r и сравнивать с равномерным кодом ($L=\lceil \log_2|A| \rceil$).
Требование к оформлению: аккуратные функции, комментарии, проверка $\Sigma p \approx 1$, вывод результатов в табличном виде.

Использование готовых библиотек оптимального кодирования запрещено.

7. Требования к отчёту

Отчёт должен содержать:
1) цель работы;
2) исходные данные;
3) краткое описание алгоритма Шеннона–Фано;
4) таблицу кодов;
5) расчёт средней длины кода;
6) сравнительный анализ;
7) выводы;
8) исходный код программы (приложение).

8. Контрольные вопросы

1. В чём состоит идея оптимального кодирования информации?
2. Почему код Шеннона–Фано относится к префиксным кодам?
3. Какую роль играют вероятности символов при построении кода?
4. Почему средняя длина кода не может быть меньше энтропии источника?
5. В чём отличие кода Шеннона–Фано от равномерного кодирования?
6. Почему код Шеннона–Фано не всегда является строго оптимальным?

9. Выводы

В выводах студент должен оценить эффективность кода Шеннона–Фано, показать его приближение к энтропийному пределу и обосновать необходимость использования алгоритма Хаффмана.