МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
Національний університет кораблебудування
імені адмірала Макарова

С.В. Блінцов, Д.О. Жук, В.С. Карпенко

ТЕОРІЯ ІНФОРМАЦІЇ

Методичні вказівки до виконання лабораторних робіт

Рекомендовано Методичною радою НУК

Миколаїв 2007

УДК 681.3

Блінцов С.В., Жук Д.О., Карпенко В.С. Теорія інформації. Методичні вказівки до виконання лабораторних робіт. – Миколаїв: НУК, 2007. – 32 с.

Кафедра комп'ютеризованих систем управління

Методичні вказівки містять методику виконання лабораторних робіт по курсам "Теорія інформації" і "Основи теорії інформації" і передбачають вивчення статистичних та ентропійних характеристик джерел інформації, методів стиснення за допомогою рівномірних і нерівномірних кодів та моделювання цих процесів за допомогою ПЕОМ. Вказівки призначені для студентів напрямів 0914 "Комп'ютеризовані системи, автоматика і управління" та 1601 "Інформаційна безпека", 0909 "Прилади точної механіки", 0915 "Спеціалізовані комп'ютерні системи".

Рецензент канд. техн. наук Турти М.В.

Згідно з наказом №110 від 25.04.07 "методичні вказівки публікуються в авторській редакції, і відповідальність за їх редагування несе автор".

ВСТУП

Даний цикл лабораторних робіт призначений для закріплення та розширення знань по основних розділах курсу "Теорія інформації" для студентів, що навчаються по спеціальності 8.091401 "Системи управління і автоматики" та курсу "Основи теорії інформації" для спеціальності 7.160101 "Захист інформації з обмеженим доступом та автоматизація її обробки в комп'ютерних системах та мережах".

Обов'язковою умовою виконання циклу є наявність персональних комп'ютерів, оснащених програмним забезпеченням Turbo або Borland Pascal та Mathcad.

В роботах передбачені вивчення статистичних та ентропійних характеристик дискретних джерел інформації, джерел марковського типу, кодування інформації з метою стиснення за допомогою рівномірних та нерівномірних кодів.

При підготовці до виконання роботи студент повинен вивчити відповідні теоретичні відомості, передбачені навчальною програмою, ознайомитись з описом лабораторної роботи. По закінченні кожної роботи студенти складають індивідуальні звіти згідно зі своїми варіантами, що містять теоретичні відомості, результати роботи з необхідними графіками, текстами програм і коментарями, висновки по роботі. Під час захисту лабораторних робіт студент повинен показати знання по відповідним розділам курсу, методам розрахунків та досліджень, виконаних у роботі.

Лабораторна робота № 1

БЕЗУМОВНІ ЙМОВІРНОСТІ ВИПАДКОВИХ ПОДІЙ

Мета роботи: дослідження апріорних імовірнісних характеристик методами цифрового статистичного моделювання.

Загальні відомості. *Випадковим* називають експеримент, результат якого не можна пророчити заздалегідь. Неможливість пророчити результат відрізняє випадкове явище від детермінованого.

Не всі випадкові явища (експерименти) можна вивчати методами теорії імовірностей, а лише ті, які можуть бути відтворені в тих самих умовах. Випадковість і хаос – не те саме. Виявляється, що у випадкових експериментах спостерігаються деякі закономірності, наприклад, властивість *статистичної стійкості*: якщо x – деяка подія, що може відбутися або не відбутися в результаті експерименту, то частка $n(x) / N$ експериментів, у яких дана подія відбулася, має тенденцію стабілізуватися з ростом загального числа експериментів N, наближаючись до деякого числа $p(x)$. Це число служить об'єктивною характеристикою "ступеня можливості" події x відбутися й називається *імовірністю* події x.

Простором елементарних результатів називається множина, що містить всі можливі результати даного випадкового експерименту, з яких в експерименті відбувається рівно один. Елементи цієї множини називають *елементарними результатами*.

Зміст домашньої підготовки

1. Вивчити опис роботи і рекомендовану літературу.

2. Скласти чорновий варіант програми, що виконує наступні дії:

в заданому місці екрану малює прямокутну або круглу область (прямокутник або коло), розмір прямокутника – $(150+10\cdot N)\times$ $\times(150+5\cdot N)$ точок, радіус кола – $(50+3\cdot N)$ точок, де N – номер варіанту;

за допомогою генератора псевдовипадкових чисел виводить на екран задану кількість точок (декілька тисяч) з випадковими координатами;

розраховує і виводить на екран експериментальну ймовірність попадання точок в область, для чого необхідно кількість точок, що попали в область, поділити на загальну кількість виведених точок;

після виведення кожної точки зберігає поточні розраховані значення експериментальної імовірності в текстовий файл "p.dat" (в кожному рядку через пробіл два числа: кількість вже виведених точок і відповідне значення імовірності) для подальшого використання в програмі Mathcad.

Порядок виконання роботи

1. Завантажити комп'ютер. В розділі диска, який дозволений для використання в студентських роботах, створити особистий каталог.

2. Увійти в середовище редактора Turbo Pascal або Borland Pascal. Завантажити і відкомпілювати програму LAB1.

3. Провести експеримент з прямокутною областю для 1000, 10000 і 100000 точок. В кожному експерименті зробити по три повторення і розрахувати середню імовірність. Розрахувати теоретичну імовірність попадання точок в область, для чого необхідно площу області поділити на площу екрану. Результати внести в таблицю наступного виду:

Кількість точок	Номер експерименту	Кількість точок, що попали в фігуру	Експериментальна імовірність	Середня експериментальна імовірність	Теоретична імовірність
1000	1				
	2				
	3				
10000	1				
	2				
	3				
100000	1				
	2				
	3				

4. Три файли з результатами підрахунку для 1000 точок завантажити в окремі змінні в середовище Mathcad (меню Insert/ Component/File Read or Write/Read from a file, формат – Text Files). Вивести відповідні графіки (в одній системі координат) залежності розрахованого значення імовірності від кількості проведених експериментів.

5. Замінити прямокутну область на коло і повторити дії, вказані в пп. 10–11.

Зміст звіту. Звіт повинен містити дві таблиці результатів розрахунку, відповідні графіки і висновки по результуючим даним.

Контрольні запитання

1. Як залежать теоретичне і експериментальне значення імовірностей від кількості проведених експериментів?
2. При яких умовах справедлива формула розрахунку теоретичного значення імовірності в експерименті, що проводиться?
3. Чи має місце залежність експериментальних значень імовірностей від форми внутрішніх областей?

Рекомендована література

Вентцель Е.С. Прикладные задачи теории вероятностей. – М.: Радио и связь, 1983.

Лабораторна робота № 2

ЗАСТОСУВАННЯ УМОВНИХ ІМОВІРНОСТЕЙ ДЛЯ ФІЛЬТРАЦІЇ НЕБАЖАНОЇ ЕЛЕКТРОННОЇ ПОШТИ

Мета роботи: отримання навичок застосування умовних імовірностей на практиці.

Загальні відомості. Умовну імовірність деякої випадкової події A відносно гіпотези H обчислюють як відношення імовірності спільного спостереження подій до безумовної імовірності гіпотези:

$p(A/H)=\frac{p(A,H)}{p(H)}$. По аналогії можна записати $p(H/A)=\frac{p(A,H)}{p(A)}$.

Якщо виразити $p(A, H)$ з другої формули і підставити в першу, отримаємо $p(A/H)=\frac{p(H/A)p(A)}{p(H)}$.

Останній вираз називають теоремою Байеса і застосовують в багатьох галузях інформаційних технологій. Наприклад, цікаве застосування теорема отримала для фільтрації небажаної елект-

ронної пошти рекламного характеру (так званого "спаму"). Розглянемо цей випадок більш докладно.

Спочатку необхідно взяти достатньо велику базу наявних електронних листів і розділити їх на дві множини – "хороша" кореспонденція (ваша реальна переписка) і "спам". Потім для кожного слова розраховується частота його появи в обох множинах. Припустимо, що словосполучення "унікальна пропозиція" зустрічається в "спам" – листах 9 разів, а в "хороших" листах – 1 раз. Тоді імовірність "спаму" відносно цього слова буде 0,9. Тобто, якщо ми в майбутньому отримаємо лист, в якому міститься це словосполучення, то з імовірністю 0,9 ми можемо віднести цей лист до "спаму", а з імовірністю 0,1 – до "хорошої" пошти. Таку умовну імовірність можна отримати для будь-якого слова з бази листів.

Переведемо задачу в терміни теорії інформації. Будемо розглядати ансамбль $Y = \{y_1, y_2\}$, де y_1 – повідомлення про те, що лист належить до "спаму", y_2 – лист належить до "хорошої" пошти.

Сформуємо ансамбль X з усіх слів, що зустрічаються в нашій базі листів. Для кожного його елементу можна розрахувати умовні ймовірності:

$$p(y_1 / x_i) = b_i / (b_i + g_i); \quad p(y_2 / x_i) = g_i / (b_i + g_i); \tag{2.1}$$

$$p(x_i / y_1) = b_i / B; \quad p(x_i / y_2) = g_i / G, \tag{2.2}$$

де b_i, g_i – кількість разів, з якою слово x_i зустрічається в базі "спаму" ("bad" mail) та в базі "хороших" листів ("good" mail), відповідно; B, G – загальна кількість слів у відповідних базах.

Тепер при надходженні нового листа можна розрахувати оцінку імовірності його відношення до "спаму" і, якщо вона більше деякого граничного значення (наприклад, більше 0,9), відфільтрувати цей лист. Для того, щоб урахувати вплив на оцінку всіх слів листа, необхідно розраховувати умовну імовірність y_1 відносно послідовності усіх слів, що містить отриманий лист:

$$p(y_1 / x^{(1)},...,x^{(n)}) = \frac{p(x^{(1)},...,x^{(n)} / y_1) \cdot p(y_1)}{p(x^{(1)},...,x^{(n)})}. \tag{2.3}$$

Будемо вважати, що слова в листі статистично не залежать одне від одного. Тоді умовна імовірність послідовності слів $x^{(1)},...,x^{(n)}$

буде визначатися добутком їх власних умовних імовірностей:

$$p(x^{(1)},...,x^{(n)}/y_1) = p(x^{(1)}/y_1)\cdots p(x^{(n)}/y_1) = \prod_{i=1}^{n} p(x^{(i)}/y_1). \quad (2.4)$$

Аналогічно, для імовірності відносно повідомлення y_2 отримаємо

$$p(x^{(1)},...,x^{(n)}/y_2) = \prod_{i=1}^{n} p(x^{(i)}/y_2).$$

Для розрахунку знаменнику в (2.3) скористаємося формулою

$$p(x_i) = \sum_{Y} p(x_i, y_j) = \sum_{Y} p(x_i/y_j)p(y_j).$$

Для послідовності повідомлень вона матиме наступний вигляд:

$$p(x^{(1)},...,x^{(n)}) =$$

$$= p(x^{(1)},...,x^{(n)}/y_1)p(y_1) + p(x^{(1)},...,x^{(n)}/y_2)p(y_2) =$$

$$= \prod_{i=1}^{n} p(x^{(i)}/y_1)p(y_1) + \prod_{i=1}^{n} p(x^{(i)}/y_2)p(y_2). \quad (2.5)$$

Будемо вважати, що ми з рівною імовірністю можемо отримати як "хороший" лист, так і "спам". Тоді $p(y_1) = p(y_2) = 0{,}5$. Це припущення робить нашу оцінку залежною лише від слів, з яких складається лист, і незалежною від частоти надходження листів з тої чи іншої групи. Врахувавши це, а також підставивши (2.4) і (2.5) в (2.3), отримаємо

$$p(y_1/x^{(1)},...,x^{(n)}) = \frac{\prod p(x^{(i)}/y_1)}{\prod p(x^{(i)}/y_1) + \prod p(x^{(i)}/y_2)}, \quad (2.6)$$

де $x^{(1)},...,x^{(n)}$ – послідовність слів, з якої складається отриманий лист; $p(x^{(i)}/y_1)$ – умовна імовірність кожного слова з цього листа.

На практиці, однак, зручніше використовувати цю формулу дещо в іншому вигляді. Знов скористаємось формулою Байеса:

$$p(x_i / y_1) = p(y_1 / x_i)\frac{p(x_i)}{p(y_1)};\ p(x_i / y_2) = p(y_2 / x_i)\frac{p(x_i)}{p(y_2)}.$$

Підставимо ці вирази в (2.6). Враховуючи, що $p(y_1) = p(y_2) = 0{,}5$, дроби $\frac{p(x_i)}{p(y_1)} = \frac{p(x_i)}{p(y_2)}$ в знаменнику можна винести за скобки і скоротити з чисельником. Крім того відзначимо, що $p(y_2 / x^{(i)}) = 1 - p(y_1 / x^{(i)})$. Тоді, остаточно, формула (2.6) розрахунку оцінки приналежності отриманого листа до "спаму" матиме вигляд:

$$p(y_1 / x^{(1)},...,x^{(n)}) = \frac{\prod p(y_1 / x^{(i)})}{\prod p(y_1 / x^{(i)}) + \prod \left(1 - p(y_1 / x^{(i)})\right)}. \quad (2.7)$$

Таким чином, необхідно лише розрахувати умовні ймовірності $p(y_1 / x_i)$ за формулою (2.1) і підставити їх в (2.7).

Зауваження 1. При початковому розрахунку умовних імовірностей слів з бази листів слід знати, що жодна імовірність не може бути встановлена в 0 чи 1, оскільки в такому випадку формула (2.7) вироджується. Якщо ми матимемо умовну імовірність відносно деякого слова рівною 1, то за формулою (2.7) ми отримаємо імовірність 1, що лист є "спамом" незалежно від імовірностей інших слів. І справа не в обмеженості формули. Вона лише показує, що якщо в листі є слово, що може зустрітися тільки в "спамі" – то лист є "спамом" на 100 %. Насправді ж, по одному слову не можна судити про весь лист. Якщо в нашій базі листів виявляться слова, що зовсім не зустрічаються в "хороших" листах, то це не значить, що вони не можуть там з'явитися взагалі. Це свідчить лише про недостатньо велику базову вибірку листів. І ймовірність таких слів необхідно встановити близькою, але не рівною одиниці (наприклад, 0,99). Аналогічна ситуація проявляється з умовною імовірністю, рівною 0. В такому випадку її слід встановлювати 0,01.

Зауваження 2. Якщо деякого слова, що зустрічається в досліджуваному листі, немає в базі, то вважатимемо його "нейтральним"

по відношенню до обох категорій листів: $p(x_i / y_1) = p(x_i / y_2) = 0{,}5$. Його можна не враховувати, оскільки його імовірності, якщо підставити їх в (2.7), все одне скоротяться.

Зміст домашньої підготовки. Вивчити опис роботи, рекомендовану літературу. Розглянути описані формули, на будь-яких текстових файлах зробити тренувальні розрахунки.

Порядок виконання роботи

1. Отримати від викладача базу слів з умовними імовірностями $p(y_1 / x_i)$, а також тестові листи, оцінку приналежності котрих до "спаму" необхідно визначити.

2. Розглянути перший лист. Сформувати таблицю, в яку виписати слова з листу (тільки ті, що містяться в базі) і відповідні ймовірності.

3. За формулою (2.7) визначити оцінку приналежності тестового листа до "спаму".

4. Повторити дії пп. 2–3 для інших наданих тестових листів.

Зміст звіту. Звіт повинен містити тексти тестових листів, таблицю умовних імовірностей слів, з яких вони складаються, результати розрахунку імовірностей приналежності тестових листів до "спаму" і висновки за результатами.

Контрольні запитання

1. В яких випадках значення умовної імовірності дорівнює нулю?

2. Як знайти імовірність пари спільних подій, якщо вони статистично залежні? Як розраховуються імовірності послідовностей повідомлень у випадку їх незалежності?

Рекомендована література

1. *Вентцель Е.С.* Прикладные задачи теории вероятностей. – М.: Радио и связь, 1983.

2. *Розанов Ю.А.* Случайные процессы. Краткий курс. – М.: Наука, 1979.

Лабораторна робота № 3

ЕКСПЕРИМЕНТАЛЬНЕ ДОСЛІДЖЕННЯ ЗАКОНУ ВЕЛИКИХ ЧИСЕЛ

Мета роботи: дослідження виконання закону великих чисел методами цифрового статистичного моделювання.

Загальні відомості. Без перебільшення можна сказати, що закон великих чисел є одними з найбільш важливих тверджень теорії імовірностей. Саме через нього теорія стикається с практикою, оскільки в ньому закладено межі застосовності теорії імовірностей.

Розглянемо закон великих чисел у формі Чебишева. Нехай X_1, ..., X_n – незалежні однаково розподілені дискретні випадкові величини, що мають кінцеве математичне очікування й дисперсію. Тоді для будь-яких додатних ε і δ знайдеться таке N, що залежить від ε і δ, що для всіх $n > N$ імовірність того, що середнє арифметичне випадкові величини X_1, ..., X_n буде відрізнятися від математичного очікування m_X кожної з випадкових величин на величину, не меншу чим ε, не перевершує δ:

$$\Pr\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - m_X\right| \geq \varepsilon\right) \leq \delta,$$

де $\delta = \dfrac{\sigma_X^2}{n\varepsilon^2}$, σ_X^2 – дисперсія випадкової величини X.

Іншими словами, закон великих чисел дає змогу розрахувати імовірність того, що після проведення n експериментів отримане середнє арифметичне буде відрізнятися від математичного очікування на величину, більшу за ε. Цю формулу можна використати для визначення імовірності деякої події з заданою похибкою.

Розглянемо двійковий ансамбль повідомлень $X = \{0, 1\}$ з імовірностями $p(1) = p$ і $p(0) = 1 - p$. Його математичне очікування

$$m_X = \sum_{i=1}^{2} x_i p(x_i) = 1 \cdot p + 0 \cdot (1 - p) = p,$$

тобто дорівнює імовірності одиниці. Середнє арифметичне $\dfrac{1}{n}\sum_{i=1}^{n} X_i$,

яке представляє собою суму нулів і одиниць, що випадали в результаті експериментів, поділену на їх кількість, фактично є відносною долею одиниць в отриманій послідовності чисел, тобто експериментально отриманою імовірністю $p_{\text{експ}}$ випадіння одиниці. Тоді згідно з законом великих чисел можна записати: $\Pr(|p_{\text{експ}} - p| \geq \varepsilon) \leq \delta$.

Таким чином, якщо стоїть задача шляхом експерименту визначити імовірність деякої події (наприклад, попадання точки в фігуру, див. лабораторну роботу № 1) з точністю ε, то по даній формулі й залежності $\delta = \dfrac{\sigma_X^2}{n\varepsilon^2}$ можна визначити кількість повторень експерименту n, які необхідно провести, щоб імовірність відхилення $p_{\text{експ}}$ від p на величину ε була менша, скажімо, за 0,001.

Зміст домашньої підготовки. Вивчити опис роботи і рекомендовану літературу. Переробити програму, розроблену в лабораторній роботі № 1 таким чином, щоб вона багатократно (від 1000 разів) повторювала виведення точок і розрахунки імовірності, обчислювала отриману імовірність відхилення $\Pr(|p_{\text{експ}} - p| \geq \varepsilon)$.

Порядок виконання роботи

1. Завантажити комп'ютер.

2. Увійти в середовище редактора Turbo Pascal або Borland Pascal. Завантажити текст програми LAB3.

3. Провести експерименти з прямокутною областю для 1000, 10000 і 100000 точок, задавши від 1000 до 5000 повторень експерименту (у залежності від кількості точок і швидкодії комп'ютера, для підвищення швидкодії треба закоментувати виведення точок на екран). Для кожної кількості точок експерименти повторити з різними значеннями ε по три рази і розрахувати середнє. Результати внести в таблицю наступного виду:

Кількість точок	Кількість повторень експерименту	ε	$\Pr(\|p_{\text{експ}} - p\| \geq \varepsilon)$ – експериментальна імовірність відхилення				δ – теоретична імовірність відхилення
			1	2	3	Середнє	
1000	5000	0,05					
	5000	0,01					
	5000	0,005					

Продовж. табл.

Кількість точок	Кількість повторень експерименту	ε	$\Pr(\lvert p_{\text{експ}} - p\rvert \geq \varepsilon)$ – експериментальна імовірність відхилення				δ – теоретична імовірність відхилення
			1	2	3	Середнє	
10000	2000	0,05					
	2000	0,01					
	2000	0,005					
100000	1000	0,05					
	1000	0,01					
	1000	0,005					

4. Замінити прямокутну область на коло і повторити експерименти, вказані в попередньому пункті, оформивши результати в таку саму таблицю.

5. В середовищі Mathcad для кожного значення e построїти графіки залежності $\delta(n)$ імовірності відхилення від кількості повторень в експерименті (по три графіки для прямокутника і для кола).

6. Позначити на графіку точки, отримані експериментально.

7. Зробити висновки щодо відповідності теоретичних залежностей і отриманих експериментально результатів.

Зміст звіту. Звіт повинен містити дві таблиці результатів експерименту, Mathcad-програму розрахунку з необхідними графіками і висновки по результуючим даним.

Контрольні запитання

1. Якою буде доля одиниці в довгій двійковій послідовності?

2. В яких межах буде лежати імовірність відхилення експериментальної і теоретичної імовірностей?

3. Чи можливо, що в довгій двійковій послідовності доля одиниць буде суттєво відрізнятися від імовірності появи одиниці? Якщо ні, то чому? Якщо так, то як зменшити імовірність такого відхилення?

Рекомендована література

Колесник В.Д., Полтырев Г.Ш. Курс теории информации. – М.: Наука. Главная редакция физ.-мат. литературы. – 1982.

Лабораторна робота № 4

ВИВЧЕННЯ ЕНТРОПІЙНИХ ХАРАКТЕРИСТИК ДИСКРЕТНИХ ДЖЕРЕЛ МАРКОВСЬКОГО ТИПУ

Мета роботи: дослідження характеристик дискретних джерел повідомлень – залежності ентропійних характеристик від часу роботи джерела, інформаційного зв'язку між повідомленнями та ін.

Загальні відомості. Дискретним джерелом, з точки зору теорії інформації, вважають будь-який пристрій U_X, який в кожну одиницю часу вибирає одне з повідомлень, що належать деякому ансамблю X. Як правило, множина X одна для кожного моменту часу, хоча в деяких випадках для кожного моменту часу може бути свій ансамбль. Фізична природа джерела не входить в коло питань теорії інформації. Джерело вважається заданим, якщо для слів будь-якої довжини, локалізованих в будь-якому часовому інтервалі, існує спосіб визначення сімейства $p(x^n)$ розподілу імовірностей, де $x^n = \{x^{(i+1)}, x^{(i+2)}, \ldots, x^{(i+n)}\}$ – слово з n повідомлень, сформоване джерелом з моменту $T = i+1$. Джерело відноситься до класу стаціонарних, якщо розподіл $p(x^n)$ не залежить від i.

Якщо джерело є простим ланцюгом Маркова, то його імовірнісні та інформаційні характеристики повністю визначені двома видами ймовірностей:

1. Абсолютними: $p(x_i^{(m)})$, де $x_i^{(m)}$ – повідомлення, яке джерело вибрало в момент часу m.

2. Умовними (перехідними): $p(x_i^{(m)} / x_j^{(l)})$, де $1 \le l \le m$.

Умовна імовірність $p(x_i^{(m)} / x_j^{(l)})$ визначає імовірність повідомлення x_i в момент m, якщо в деякий попередній момент l джерело вибрало повідомлення x_j. Зазвичай ці умовні ймовірності називають імовірностями переходу від повідомлення x_j до повідомлення x_i за $(m - l)$ кроків.

Очевидно, що введені ймовірності є додатними і задовольняють умові нормування:

$$1.\sum_{i=1}^{L} p(x_i) = 1; \quad 2.\sum_{i=1}^{L} p(x_i / x_j) = 1, \; p(x_i / x_j) \ge 0, \quad i = 1,\ldots,L.$$

На основі теореми повної імовірності для безумовної імовірності повідомлення x_i в момент часу m можна записати:

$$p(x_i^{(m)}) = \sum_{j=1}^{L} p(x_i^{(m)}, x_j^{(l)}) = \sum_{j=1}^{L} p(x_j^{(l)}) p(x_i^{(m)} / x_j^{(l)}), \quad i = 1,...,L.$$

Використовуючи формулу повної імовірності й визначення ланцюга Маркова можна пересвідчитися, що

$$p(x_k^{(c)} / x_j^{(a)}) = \sum_{i=1}^{L} p(x_k^{(c)} / x_i^{(b)}) p(x_i^{(b)} / x_j^{(a)}),$$

$$j,k = 1,...,L; \; 0 \le a < b < c.$$

Наприклад, для ансамблю з трьох повідомлень маємо:

$$p(x_2^{(3)} / x_1^{(1)}) = p(x_2^{(3)} / x_1^{(2)}) p(x_1^{(2)} / x_1^{(1)}) + p(x_2^{(3)} / x_2^{(2)}) p(x_2^{(2)} / x_1^{(1)}) +$$

$$+ p(x_2^{(3)} / x_3^{(2)}) p(x_3^{(2)} / x_1^{(1)}).$$

В матричному вигляді (для стислості запишемо для ансамблю з двох повідомлень):

$$p(x/x)^2 = \begin{bmatrix} p(x_1^{(k+1)} / x_1^{(k)}) & p(x_2^{(k+1)} / x_1^{(k)}) \\ p(x_1^{(k+1)} / x_2^{(k)}) & p(x_2^{(k+1)} / x_2^{(k)}) \end{bmatrix} \times$$

$$\times \begin{bmatrix} p(x_1^{(k+2)} / x_1^{(k+1)}) & p(x_2^{(k+2)} / x_1^{(k+1)}) \\ p(x_1^{(k+2)} / x_2^{(k+1)}) & p(x_2^{(k+2)} / x_2^{(k+1)}) \end{bmatrix} =$$

$$= \begin{bmatrix} p(x_1^{(k+1)} / x_1^{(k)}) & p(x_2^{(k+1)} / x_1^{(k)}) \\ p(x_1^{(k+1)} / x_2^{(k)}) & p(x_2^{(k+1)} / x_2^{(k)}) \end{bmatrix} = p(x_i^{(k+2)} / x_j^{(k)}).$$

Якщо матрицю умовних імовірностей піднести до третього степеня, тобто $p(x_i^{(k+2)} / x_j^{(k)})$ помножити ще раз на $p(x / x)$, то отримаємо перехідні ймовірності повідомлень, що розташовані вже через три кроки одне від одного. Аналогічно можна отримати перехідні ймовірності для повідомлень на будь-якій відстані.

При обмеженій довжині слова середня кількість інформації, що приходиться на одне повідомлення джерела (ентропія на повідомлення), визначається таким чином:

$$H(X/X^{\infty}) = \lim_{n\to\infty} H_n(X) = \lim_{n\to\infty} \frac{H(X^n)}{n}. \qquad (3.1)$$

Оскільки $H(X^n) = H(X^{n-1}X) = H(X^{n-1}) + H(X/X^{n-1})$, то приріст інформації за одну одиницю часу складе $\Delta H_n = H(X/X^{n-1})$. Для марковського ланцюга першого порядку умовна ентропія залежить лише від останнього моменту часу: $\Delta H_n = H(X/X)$. Таким чином, ентропію на повідомлення можна розрахувати безпосередньо за формулою (3.1), попередньо визначивши ймовірності всіх послідовностей ансамблю X^n, або скориставшись властивістю адитивності ентропії:

$$H(X^n) = H(X_1) + H(X_2/X_1) + \ldots + H(X_n/X_{n-1}).$$

Інформаційний зв'язок між першим $x^{(1)}$ і останнім $x^{(n)}$ повідомленнями визначається взаємною інформацією:

$$I(X_1; X_n) = H(X_n) - H(X_n/X_1).$$

Зміст домашньої підготовки. Вивчити опис роботи і рекомендовану літературу. Підготувати попередній варіант програми розрахунків в системі Mathcad.

Порядок виконання роботи

1. Завантажити комп'ютер. Запустити Mathcad.

2. Ввести допоміжні формули і початкові дані згідно свого номеру варіанту.

Розглянути дискретне джерело повідомлень, що в кожний момент часу випадково вибирає один елемент ансамблю $X = \{x_1, \ldots, x_5\}$ згідно з розподілом імовірностей. В початковий момент часу ($k = 1$) імовірності дорівнюють

$$p = [p(x_1^{(1)})\ p(x_2^{(1)})\ p(x_3^{(1)})\ p(x_4^{(1)})\ p(x_5^{(1)})] = [1-0{,}01N\ \ 0{,}01N\ \ 0\ \ 0\ \ 0],$$

де N – номер варіанту.

Умовні перехідні ймовірності:

$$p(x_i^{(k+1)}/x_j^{(k)}) = \begin{bmatrix} p(x_1^{(k+1)}/x_1^{(k)}) & \dots & p(x_5^{(k+1)}/x_1^{(k)}) \\ p(x_1^{(k+1)}/x_2^{(k)}) & \dots & \dots \\ p(x_1^{(k+1)}/x_3^{(k)}) & \dots & \dots \\ p(x_1^{(k+1)}/x_4^{(k)}) & \dots & \dots \\ p(x_1^{(k+1)}/x_5^{(k)}) & \dots & p(x_5^{(k+1)}/x_5^{(k)}) \end{bmatrix} =$$

$$= \begin{bmatrix} 0 & 0{,}4 & 0{,}6 & 0 & 0 \\ 0 & 0{,}5 & 0{,}2 & 0 & 0{,}3 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0{,}5 & 0 & 0 & 0{,}5 \\ 0{,}1 & 0 & 0 & 0 & 0{,}9 \end{bmatrix}.$$

3. Виконати розрахунок абсолютних імовірностей і ентропії ансамблю в кожний момент часу. Вибрати тривалість роботи джерела і по графіку візуально оцінити час переходу до стаціонарного режиму.

4. Виконати розрахунки середньої кількості інформації на повідомлення (двома способами) і вивести на екран її залежність від часу.

5. Виконати розрахунки умовної ентропії та взаємної інформації між ансамблями, розділеними інтервалом в один, два і більше тактів.

Зміст звіту. Звіт повинен початкові дані, текст Mathcad-програми, результати розрахунків і необхідні графіки.

Контрольні запитання

1. Як залежить ентропія від розміру ансамблю джерела?
2. Як змінюється характер перехідного процесу із змінами початкових умов?
3. Який зв'язок між продуктивністю джерела і кількістю змін повідомлень?
4. Які характерні риси мають зміни інформаційного зв'язку між першим і останнім повідомленнями слова при змінах його розмірів?

Рекомендована література

1. *Колесник В.Д., Полтырев Г.Ш.* Курс теории информации. –

М.: Наука. Главная редакция физико-математической литературы. – 1982.

2. *Тихонов В.И.* Марковские процессы. – М.: Советское Радио, 1977.

Лабораторна робота № 5

ВИСОКОІМОВІРНІ МНОЖИНИ ПОВІДОМЛЕНЬ ДИСКРЕТНИХ ДЖЕРЕЛ БЕЗ ПАМ'ЯТІ

Мета роботи: вивчення принципів рівномірного кодування з використанням теореми про високоімовірні множини.

Загальні відомості. Ідея про те, що загальний випадок нерівноймовірних можливостей (станів) асимптотично зводиться до випадку рівноймовірних, лежить в основі теорії кодування у відсутність перешкод. Ця ідея належить Л. Больцману, який одним з перших вивів формулу для ентропії. К. Шеннон відродив цю ідею і широко використав для отримання нових результатів.

Візьмемо набір незалежних реалізацій $x' = (x^{(1)}, x^{(2)}, ..., x^{(n)})$ випадкового повідомлення хі, що приймає одне з значень ансамблю X. Очевидно, що число всіх різних наборів x' дорівнює L^n, де L – кількість повідомлень в ансамблі X. Згідно з теоремою про високоімовірні множини, повну множину X^n всіх послідовностей можна розділити на дві підмножини – $T_n(\varepsilon)$ і $\overline{T}_n(\varepsilon)$. До першої відносяться послідовності, в яких середня кількість інформації на повідомлення близька до ентропії $H(X)$ ансамблю X:

$$T_n(\varepsilon) = \{x' : H(X) - \varepsilon \le \frac{1}{n} I(x') \le H(X) + \varepsilon\},$$

де $I(x') = -\log p(x')$ – власна інформація послідовності x', n – довжина послідовності, ε – деяке додатне число.

Послідовності з множини $T_n(\varepsilon)$ називаються високоімовірними завдяки двом своїм властивостям:

сумарна імовірність їх появи може бути зроблена як завгодно близькою до одиниці із збільшенням n:

$$\Pr(x' \in T_n(\varepsilon)) \ge 1 - \delta,$$

де $\delta = \dfrac{\sigma_I^2}{n\varepsilon^2}$, σ_I^2 – дисперсія кількості власної інформації в повідомленнях ансамблю X;

імовірності цих послідовностей близькі одна до одної і лежать в досить вузьких межах:

$$2^{-n[H(X)+\varepsilon]} \le p(x') \le 2^{-n[H(X)-\varepsilon]}.$$

Множина $\overline{T}_n(\varepsilon)$ – це доповнення множини $T_n(\varepsilon)$ до X^n. Очевидно, що послідовності з цієї множини майже не появляються, оскільки сумарна їх імовірність не більша за число δ, яке може бути зроблене як завгодно малим із збільшенням n.

Слід також відзначити, що множина $T_n(\varepsilon)$ може становити дуже малу частку по числу елементів від множини X^n. Дійсно, якщо позначити число елементів в X через L, то кількість усіх послідовностей $\left|X^n\right| = L^n = 2^{n\log L}$. Згідно з теоремою про високоімовірні множини, кількість послідовностей $|T_n(\varepsilon)|$ у множині $T_n(\varepsilon)$ визначається нерівностями $(1-\delta)2^{n[H(X)-\varepsilon]} \le |T_n(\varepsilon)| \le 2^{n[H(X)+\varepsilon]}$.

Тоді частка α множини $T_n(\varepsilon)$ в X^n дорівнює:

$$\alpha = \frac{|T_n(\varepsilon)|}{\left|X^n\right|} \le 2^{-n[\log L - H(X)-\varepsilon]}.$$

Якщо $\log L - H(X) - \varepsilon = k > 0$, тобто якщо ентропія джерела строго менша, ніж log L, і ε достатньо мале, то α убуває до нуля при збільшенні n як 2^{-kn}, тобто як степенева функція.

Описані властивості високоімовірної множини використовують для рівномірного кодування. Послідовності з множини $T_n(\varepsilon)$ кодують однозначно, всі інші – будь-яким одним кодовим словом. Таким чином, кількість кодових слів зменшується з $|X^n|$ до $|T_n(\varepsilon)|$, що в α разів менше. При цьому імовірність похибки при декодуванні не буде перевищувати δ, тобто імовірності появи повідомлення з $\overline{T}_n(\varepsilon)$, і може бути зробленою як завгодно малою.

Для кодування необхідно стільки кодових слів, скільки послідовностей входить в $T_n(\varepsilon)$. Тоді довжина кодового слова (в бітах) має бути $m = \log_2 T_n(\varepsilon)$. Оскільки $|T_n(\varepsilon)| \le 2^{n[H(X)+\varepsilon]}$, то

$$m \le n[H(X)+\varepsilon].$$

При цьому кожне кодове слово кодує послідовність з n повідомлень. Очевидно, що на одне повідомлення при кодуванні буде витрачатись не більше $H(X)+\varepsilon$ біт.

Зазначимо, що $\varepsilon \to 0$ при $n \to \infty$. Таким чином, для наближення до теоретично можливого мінімального значення кількості біт на повідомлення, рівного ентропії ансамблю $H(X)$, і отримання максимального стиснення необхідно кодувати якомога довші послідовності.

Зауваження. Оскільки файли мають обмежену довжину, то вони, зазвичай, містять далеко не всі можливі високоімовірні послідовності, а лише деяку частку з них. Відповідно, фактична кількість кодових слів і довжина кодового слова для їх кодування є значно меншою. Але це не заважає досліджувати вплив похибки при появі низькоімовірних послідовностей на результати декодування.

Зміст домашньої підготовки

1. Вивчити опис лабораторної роботи, відповісти на контрольні запитання.

2. Підготувати попередні варіанти програм, використовуючи опис, наведений нижче.

Програма "stats.pas" повинна виконувати наступні дії:

отримати ім'я файлу для досліджень з командного рядку чи з клавіатури (або задати його безпосередньо в тексті програми);

розрахувати розподіл імовірностей в досліджуваному файлі для ансамблю повідомлень $X = \{x_0, x_1, ..., x_{255}\}$, елементами якого є числа від 0 до 255 (тобто 1-байтні числа), а імовірності розраховуються за формулою

$$p(x_i) = n_i / N,$$

де n_i – кількість разів, з якою байт хі зустрічається в досліджуваному файлі, $i = 0, ..., 255$; N – довжина файлу в байтах;

записати отримані ймовірності в бінарний файл даних "p.dat", а також в текстовий файл "p.txt" в наступному форматі:

0 – 0.015433
1 – 0.004249
...
255 – 0.007127,

де в кожному рядку перше число – це байт від 0 до 255, друге – відповідна імовірність з точністю 6 знаків після коми;

розрахувати ентропію ансамблю X і вивести її значення на екран, а також в останній рядок файлів "p.dat" і "p.txt".

Програма " coder.pas" повинна виконувати наступні дії:

зчитати імовірності та ентропію ансамблю X з файлу "p.dat";

побудувати кодову таблицю, для чого необхідно: зчитувати з досліджуваного файлу послідовності довжини n; перевіряти їх на належність до високоімовірної множини; перевіряти, чи не зустрічається вже ця послідовність в кодовій таблиці, якщо ні – записувати послідовність в файл кодової таблиці "cod.dat";

закодувати досліджуваний файл за допомогою кодової таблиці, що знаходиться в файлі "cod.dat", і записати результат кодування в бінарний файл "packed.dat";

декодувати файл "packed.dat" за допомогою кодової таблиці з файлу "cod.dat" і записати результат в бінарний файл "unpacked.bmp";

переписати заголовок вхідного bmp-файлу в декодований файл і зберегти у файлі "unpackd1.bmp", щоб можна було переглянути декодований графічний файл у разі пошкодження заголовку при кодуванні з великою імовірністю похибки.

Порядок виконання роботи

1. Завантажити комп'ютер. Запустити Turbo Pascal.
2. Завантажити і скомпілювати текст програми "stats.pas".
3. Завантажити і скомпілювати текст програми "coder.pas".
4. Обрати файл для кодування/декодування (будь-який файл в форматі .bmp) та скопіювати його в робочу директорію ("..\Data").
5. За допомогою програми "stats.pas" виконати розрахунки імовірнісних характеристик обраного файлу.
6. Для декількох значень довжини послідовності n і ймовірності похибки δ виконати програму "coder.pas" (прийняти $\delta = 0{,}001; 0{,}01; 0{,}02; 0{,}05$; $n = (200 + N); (500 + N); (1000 + N)$, де N – номер варіанту, скомбінувати між собою всі значення δ і n).

Для кожного випадку розрахувати (за допомогою калькулятора системи Windows) характеристики коду – кількість високоімовірних послідовностей $|T_n(\varepsilon)|$, кількість усіх послідовностей $|X^n|$, долю α високоімовірних послідовностей, довжину m кодового слова відповідного рівномірного коду.

7. Зробити висновки щодо ступеня стиснення і викривлень, які відбуваються з кодованим файлом в залежності від n і δ.

Зміст звіту. Звіт повинен містити таблицю імовірностей файлу та його ентропію, декілька прикладів декодованих файлів (з викривленнями і без таких) з відповідними розрахунками, що робить програма, а також розраховані для кожного випадку характеристики коду, висновки по роботі.

Контрольні запитання

1. Яка кількість слів двійкового ансамблю, що складені з n символів, утворить множину $T_n(\varepsilon)$ при імовірності $p = 0{,}1$?

2. Як змінюється розмір інтервалу, який включає типові послідовності, при змінах довжини слова і постійній імовірності p?

3. Які слова джерела повинні увійти в множину $T_n(\varepsilon)$, якщо $p = 0{,}5$?

4. При яких умовах кількість власної інформації слова, що складається з $n > 1$ символів, виявиться рівною одному біту?

Рекомендована література

1. *Дмитриев В.И.* Прикладная теория информации. – М.: Высшая Школа, 1989.

2. *Стратонович Р.Л.* Теория информации. – М.: Советское Радио, 1975.

Лабораторна робота № 6

ОПТИМАЛЬНЕ НЕРІВНОМІРНЕ КОДУВАННЯ МЕТОДОМ ХАФФМАНА

Мета роботи: вивчення властивостей нерівномірних кодів і методів їх отримання програмним шляхом.

Загальні відомості. Задачею ефективного кодування є такий спосіб обробки дискретного сигналу, при якому середня кількість одиниць інформації на елементарне повідомлення настільки мала, наскільки це принципово можливо. При цьому потрібно так записати повідомлення, щоб по запису можна було відновити їх без втрат і помилок. К. Шенноном (Claude Elwood Shannon) було доведено (1948 р.), що завжди можна побудувати таку систему ефективного кодування дискретного джерела, при якій середня кількість двійкових знаків на символ джерела як завгодно близька, але не менша за ентропію джерела на повідомлення.

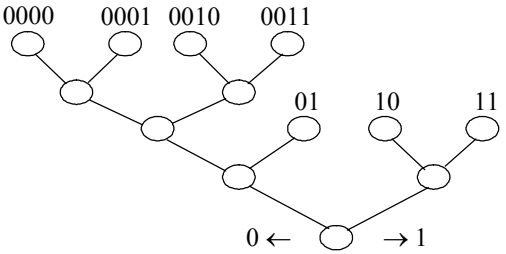
В системі нерівномірного кодування кодові слова мають різну довжину m_i, яка залежить від повідомлення x_i, що кодується. Такий код характеризують середньою довжиною кодових слів:

$$\overline{m}(X) = \sum_{x_i \in X} m_i p(x_i).$$

Код вважається оптимальним, якщо $\overline{m}(X) \to H(X)$. Якщо в повідомленнях джерела кодують не кожний символ, а блоки $x' \in X^n$, то оптимізаційна задача формулюється таким чином: мінімізувати $\overline{m}(X^n) \to H(X^n)$, де n – довжина блоку x' (слова). Доведена теорема, згідно з якою середня довжина оптимального нерівномірного коду міститься в інтервалі

$$H(X^n) \le \overline{m}(X^n) \le H(X^n) + 1.$$

При невеликих об'ємах нерівномірний код може наочно представлятися у вигляді кодових дерев, як показано на наступному рисунку.



Вузли дерева, що відходять від кореня на i ребер, утворюють ярус порядку i. Порядком вузла називають номер його ярусу. Порядком дерева називають максимальний з порядків його вузлів. Вузол, з якого не виходить жодного ребра, називається кінцевим. Якщо кодові слова відповідають тільки кінцевим вузлам, то код є префіксним. Для існування префіксного двійкового коду необхідно і достатньо виконання нерівності

$$\sum_{x_i \in X} 2^{-m_i} \le 1,$$

яку називають нерівністю Крафта. У 1952 р. Д. Хаффман (David A. Huffman) знайшов нескладний спосіб оптимального кодуван-

ня, який забезпечує отримання мінімальної середньої довжини кодових слів з усіх можливих для заданого джерела. Ключовим моментом в алгоритмі побудови оптимального коду є процес утворення кодових слів за наступним принципом: повідомленню з мінімальною імовірністю ставиться у відповідність найдовше кодове слово.

Якщо двійковий код оптимальний, то кодове дерево повинно мати наступну обов'язкову властивість. З кожного внутрішнього вузла, а також з кореня, завжди повинні вийти 2 ребра. В іншому випадку єдине ребро можна було б "стиснути" в цей неповний вузол і тим самим скоротити середню довжину кодового слова. Процес побудови кодового дерева починається від кінцевих вершин, яким на першому етапі привласнюються номери повідомлень, що кодуються разом з їх імовірностями. Другий етап, що складається з повторюваних операцій, виконується доти, поки не буде отримано корінь кодового дерева. На початку цього етапу відшукують пару вершин v_i, v_j, які не мають вхідних ребер і мають мінімальні імовірності $p(v_i)$ і $p(v_j)$. Потім додають нову вершину v_k і з'єднують її парою ребер з v_i і v_j. Імовірність $p(v_k)$ знаходять як $p(v_k) = p(v_i) + p(v_j)$.

Після завершення цього кроку кількість вузлів, що не мають вхідних ребер, зменшується на одиницю. Якщо таких вузлів залишиться два, то вузол, що знову додається, буде коренем.

Ефективність кодування визначається надмірністю. Її мірою служить величина D, що показує, наскільки добре використовуються знаки даного джерела:

$$D = 1 - \frac{H(X)}{H_{\max}(X)},$$

де $H(X)$ – ентропія джерела, яке вибирає повідомлення з ансамблю з L елементів; $H_{\max}(X) = \log_2 L$ – максимально можлива ентропія такого ансамблю.

Якщо надмірність джерела дорівнює нулю, то створювані ним повідомлення оптимальні в сенсі найбільшої кількості інформації, що передається. Для передачі певної кількості інформації I при відсутності перешкод в цьому випадку необхідно $k_1 = I / H_{\max}(X)$ знаків.

Оскільки ентропія повідомлень, створюваних реальним джерелом, що має надмірність, менше максимальної, то для передачі

тієї ж кількості інформації I знаків потрібно більше, а саме: $k_2 = I / H(X) > k_1$. Тому говорять також про надмірність знаків у повідомленні або просто про надмірність повідомлення, характеризуючи її тим же самим параметром D:

$$D = 1 - \frac{k_1}{k_2} = 1 - \frac{H(X)}{H_{\max}(X)}.$$

Зміст домашньої підготовки

1. Вивчити опис роботи, рекомендовану літературу.

2. Підготувати попередній варіант програми, використовуючи опис, наведений нижче.

Програма "Huffman.pas" повинна виконувати наступні дії:

розрахувати імовірності й ентропію вхідного файлу (по аналогії з лабораторною роботою № 5);

побудувати відповідний оптимальний нерівномірний код за методом Хаффмана і записати його в текстовий файл "haffcode.dat", де перший рядок повинен містити кодове слово, що відповідає числу $x_0 = 0$, другий – числу $x_1 = 1$ і т.д. до 256-го рядка, який відповідає байту 255;

записати нерівномірний код в текстовий файл "Report.txt" в наступному форматі:

0 – 0.015433 – 00011010

...

255 – 0.007127 – 10010010111,

де в кожному рядку перше число – це байт від 0 до 255, друге – його імовірність, третє – відповідне кодове слово;

розрахувати й вивести на екран теоретичну середню довжину кодового слова, що обчислюється за формулою $\overline{m}(X) = \sum_{i=0}^{255} m_i p(x_i)$,

де m_i – довжина кодового слова, яким кодується елемент x_i;

закодувати файл за допомогою нерівномірного коду і записати результат кодування у вигляді послідовності символів "0" і "1" в текстовий файл "packed.txt", а також у вигляді послідовності біт в бінарний файл "packed.dat";

розрахувати й вивести на екран ентропію закодованого файлу (для цього знов необхідно розрахувати ймовірності появи кожного байта);

розкодувати файл "packed.txt" за допомогою нерівномірного коду і записати результат в бінарний файл "unpacked.bmp".

Порядок виконання роботи

1. Завантажити комп'ютер. Запустити Turbo Pascal.

2. Завантажити і скомпілювати текст програми "Huffman.pas".

3. Обрати файл для кодування/декодування, наприклад, будь-який файл .bmp або .txt з папки Windows та скопіювати його в робочу директорію програми.

4. Виконати програму "Huffman.pas". Пересвідчитися, що декодований файл "unpacked.bmp" співпадає з оригіналом. Записати в звіт довжину вхідного та закодованого файлів, а також виведені на екран результати роботи програми.

5. Розрахувати фактично отриману середню довжину кодового слова, яка обчислюється за формулою

$$m_{\text{ф}} = N_c / N,$$

де N_c – довжина закодованої послідовності (в бітах); N – кількість закодованих повідомлень, тобто довжина вхідного файлу в байтах.

6. Розрахувати надмірність вхідного файлу і надмірність закодованої послідовності.

7. Розрахувати коефіцієнт стиснення інформації, що дорівнює відношенню довжини закодованої послідовності (в бітах) до довжини вхідного файлу (також в бітах).

8. Розглянути ансамбль повідомлень X^2, елементами якого є двобайтні числа від 0 до 65535. Переробити програму для роботи з послідовністю з двох байт. Знов виконати пп.4-7 для кодування того ж самого файлу, але вже не побайтно, а послідовностями по 2 байта.

9. Повторити кодування-декодування і розрахунки (пп.4-8) ще для двох будь-яких файлів.

Зміст звіту. Звіт повинен містити вхідні дані – імена та розміри досліджуваних файлів, а також отримані характеристики кодування ансамблей X і X^2 для кожного файлу, зведені в таблицю, що містить розмір початкового і закодованого файлів, ентропію і надмірність ансамблю, теоретичну і фактичну середню довжину кодового слова, ентропію і надмірність закодованої послідовності, коефіцієнт стиснення інформації. У висновках необхідно проаналізувати ефективність нерівномірного кодування, маючи на увазі

основну його мету – зменшення надмірності до нуля, показати шлях для такого зменшення.

Контрольні запитання

1. Наведіть необхідну і достатню умову існування коду з властивістю однозначного декодування.
2. Назвіть недоліки оптимальних нерівномірних кодів.
3. Назвіть шлях отримання нерівномірного коду, в якому витрата бітів на одне повідомлення буде якомога близькою до мінімально можливої.
4. Чи можливо декодувати залишок повідомлення, якщо початок передачі невідомий?

Рекомендована література

1. *Колесник В.Д., Полтырев Г.Ш.* Курс теории информации. – М.: Наука. Главная редакция физико-математической литературы. – 1982.
2. *Стратонович Р.Л.* Теория информации. – М.: Советское Радио, 1975.

Лабораторна робота № 7

КОДУВАННЯ ДВІЙКОВИХ ДЖЕРЕЛ БЕЗ ПАМ'ЯТІ МЕТОДОМ "ДОВЖИН СЕРІЙ"

Мета роботи: вивчення методу перекодування двійкових послідовностей для потреб стиснення інформації.

Загальні відомості. На відміну від поширених методів, метод, що отримав назву "кодування довжин серій" (КДС) не пов'язаний з групуванням послідовності повідомлень на слова фіксованої довжини. З цього методу слідує, що крім задачі вибору оптимального коду для даної множини особливий інтерес викликає також задача вибору *хорошої множини* повідомлень, що кодуються. Нехай джерело породжує послідовність двійкових незалежних символів "0" і "1", причому імовірність появи одиниці $p < 1/2$.

Ця послідовність розбивається на сегменти вигляду <1>, <01>, <001>, …, <0…01>, <0…000>, де довжини сегментів відповідно дорівнюють 1, 2, …, N. Кількість сегментів дорівнює $N + 1$ (два останніх сегменти <0…01> і <0…000> мають однакову довжи-

ну). Якщо відомий початок передачі повідомлень, то групування на сегменти однозначне. Далі сегменти розглядаються як повідомлення деякого джерела; імовірність появи i-го сегмента дорівнює pq^{i-1}, $i = 1, 2, \ldots, N$, $q = 1 - p$. Імовірність появи сегмента <000…00> рівна q^N. Ці сегменти кодуються нерівномірним кодом так, що останньому сегменту зіставляється слово одиничного розміру, а іншим – слова довжини $M + 1$. Значення M вибирають з інтервалу $N \le 2^M < N + 1$.

Середня кількість інформації, яку буде мати сегмент, очевидно, дорівнює $H_N(X) = -\sum_{i=1}^{N} pq^{i-1} \log\left(pq^{i-1}\right) - q^N \log\left(q^N\right)$, а середня довжина сегменту має вигляд $n_{\text{ср}}(N) = \dfrac{1-q^N}{p}$.

Прийнятий метод кодування забезпечує наступне значення середньої довжини кодового слова: $m_{\text{ср}}(N) = 1 + M(1-q^N)$. Щоб кодування не мало втрат інформації, необхідно забезпечити наступну нерівність: $H_N(X) \le m_{\text{ср}}(N)$. Швидкість кодування, тобто кількість біт, що витрачається на кодування одного повідомлення, можна визначити за формулою $R(N) = m_{\text{ср}}(N) / n_{\text{ср}}(N)$.

Характер поведінки нижньої межі інтервалу, ентропії ансамблю сегментів та їх середнього розміру при імовірності $p = 0{,}2$ показаний на рис. 7.1,***а***.

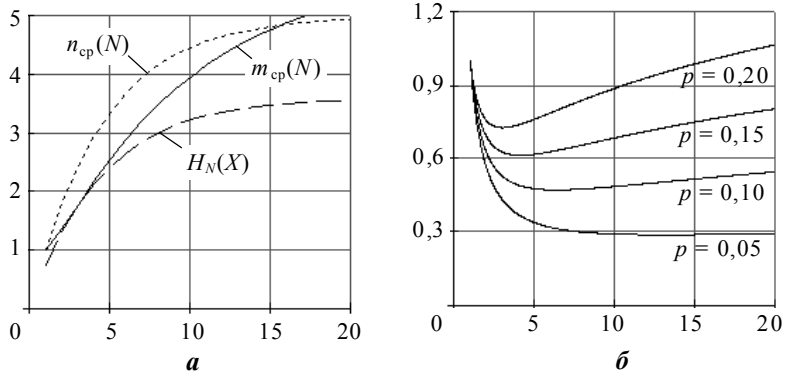


Рис. 7.1. Залежності ентропії ансамблю сегментів, середньої довжини кодових слів і середньої довжини сегментів (***а***), а також швидкості кодування (***б***) від максимальної довжини сегменту N

Як видно з графіка, ентропія і середній розмір сегмента асимптотично наближаються до постійних значень при необмеженому збільшенні N. Середній розмір кодових слів, починаючи з певних значень N починає обганяти спочатку значення ентропії, а потім і середній розмір сегмента. Така поведінка свідчить про існування деякого оптимального значення N при фіксованій імовірності p. Це очевидно також з рис. 7.1.*б*, де показана швидкість кодування для різних значень p.

Зміст домашньої підготовки. Вивчити опис лабораторної роботи. Підготувати чорновий варіант Mathcad-програми розрахунків оптимального значення кількості сегментів при фіксованих значеннях імовірності p.

Порядок виконання роботи

1. Завантажити комп'ютер, запустити систему Mathcad. Ввести програму розрахунків оптимального значення кількості сегментів.

2. Для заданого (згідно з варіантом) значення імовірності p побудувати залежності ентропії, середньої довжини кодових слів і середньої довжини сегментів, а також швидкості кодування. Визначити по графікам оптимальну довжину кодових слів M і кількість сегментів N.

3. Виконати розрахунок середньої довжини сегментів й ентропії їх ансамблю.

4. Виконати розрахунки середньої довжини кодових слів.

5. Порівняти отриману швидкість кодування з ентропією початкового двійкового ансамблю X.

Зміст звіту. Звіт повинен містити вхідні дані, текст Mathcad-програми, кількісні характеристики коду та висновки щодо оптимальності отриманого коду.

Контрольні запитання

1. Як має змінюватись значення оптимальної довжини коду при зміні ймовірності p?

2. Яким умовам повинно відповідати значення ймовірності p, щоб метод Хаффмана сформував код, що відповідає структурі КДС?

3. Викривлення яких кодових символів у словах КДС може призвести до трекових помилок при декодуванні?

Рекомендована література

1. *Дмитриев В.И.* Прикладная теория информации. – М.: Высшая Школа, 1989.

2. *Колесник В.Д., Полтырев Г.Ш.* Курс теории информации. – М.: Наука. Главная редакция физико-математической литературы. – 1982.

ЗМІСТ

Навчальне видання

БЛІНЦОВ Сергій Володимирович
ЖУК Дмитро Олександрович
КАРПЕНКО Володимир Сергійович

ТЕОРІЯ ІНФОРМАЦІЇ

Методичні вказівки до виконання лабораторних робіт

(*українською мовою*)

Комп'ютерна правка та верстка *Н.В. Ялова*
Коректор *М.О. Паненко*

Свідоцтво про внесення суб'єкта видавничої справи до Державного реєстру видавців, виготівників і розповсюджувачів видавничої продукції ДК № 2506 від 25.05.2006 р.

Підписано до друку 18.07.07. Папір офсетний. Формат 60×84/16.
Друк офсетний. Гарнітура "Таймс". Ум. друк. арк. 1,8. Обл.-вид. арк. 1,9.
Тираж 300 прим. Вид. № 23. Зам. № 184. Ціна договірна

Видавець і виготівник Національний університет кораблебудування,
54002, м. Миколаїв, вул. Скороходова, 5