ИСПОЛЬЗОВАНИЕ АЛГОРИТМОВ НА ОСНОВЕ ЛИНЕЙНОГО КОДА В СХЕМАХ ОПРЕДЕЛЕНИЯ И КОРРЕКЦИИ ОШИБОК

Гарнышев И.Н.[1], Казанцев С.В.[2], Мальков Р.Ю.[3], Семенов И.Д.[4], Юдин С.В.[5]

[1]Гарнышев Игорь Николаевич - сетевой инженер,
Отдел администрирования сетей передачи данных,
Тинькофф Банк;
[2]Казанцев Сергей Владимирович - главный инженер,
Департамент сетей передачи данных,
Сбербанк;
[3]Мальков Роман Юрьевич – эксперт
Центр компетенций по облачным решениям,
Техносерв,
г. Москва;
[4]Семенов Иван Дмитриевич - старший инженер,
Департамент сетей передачи данных,
Servers.com Лимассол, Кипр;
[5]Юдин Степан Вячеславович - администратор сети,
Департамент технического обеспечения и развития инфраструктуры информационных систем,
Спортмастер, г. Москва

Аннотация: *в статье проведен анализ принципов моделирования дискретного канала без памяти, в частности линейного блочного кодирования. Построена обобщенная схема определения пропускной способности дискретного канала. Разработан математический аппарат для определения блочного кода для дискретного канала. Предложена математическая модель декодирования линейного кода, который потенциально может содержать в себе ошибки. В результате проведенного исследования разработан универсальный алгоритм определения вероятности ошибки при декодировании блочного кода в случае передачи данных на двоичном симметричном канале.*

Ключевые слова: *дискретный канал без памяти, блочный код, матрица генерации, скорость линейного кода, общее количество информации, решающее правило по методу максимального правдоподобия, двоичный симметричный канал.*

Введение

На сегодняшний день модель информационного канала предлагается в качестве базовой модели канала связи, в рамках которого передаваемое сообщение рассматривается как вход, а его воспроизведение приемником — выходом. Та же концепция может быть использована как для моделирования системы хранения информации, причем в рамках такого подхода вход и выход рассматриваются как элементы системы, которые разделены во времени. Универсальная модель может включать физическую среду передачи данных: как непрерывную, так и квантованную. В процессе передачи и приема данных используются конечные алфавиты, которые представляют собой цифровую реализацию процессов передачи информации, независимо от их физической реализации.

Анализ последних исследований и публикаций в данной области позволил обобщить представления о принципах моделирования дискретного канала без памяти [1-3] и обозначить приоритет линейного блочного кодирования [4-6]. Показана актуальность задачи построения математической модели декодирования линейного кода с потенциальными ошибками [8-10], а также использования кода Хэмминга [11-13]. Также были рассмотрены исследования, направленные на определение вероятности ошибки при декодировании блочного кода, подготовленного для передачи данных через двоичный симметричный канал [14-17].

Целью работы стало построение комплексной методологии по работе с линейным кодом в помехоустойчивых системах передачи оцифрованных данных и специализация данного подхода для системы двоичного симметричного канала.

1. Принципы моделирования дискретного канала без памяти

Дискретный канал без памяти может быть определен как канал передачи данных, для которого вход и выход представляют собой последовательности символов, причем текущее значение выхода зависит только от текущего значения входа [1-3]. В основу математической модели дискретного канала без памяти должны быть положены следующие составные компоненты (рис. 1):

- пара переменных (X, Y);
- конечные алфавиты $A_x \in \{X_i\}$, где $i \in [1; I]$ и $A_y \in \{X_j\}$, где $j \in [1; J]$, которые определяют наборы возможных значений переменных X и Y, соответственно;
- условная вероятность $P(y_j|x_i)$ как функция, которая определяет связь между переменными X и Y;

- общее количество информации $I(X,Y)$, функция, которая определяет выбор входных символов в соответствии с необходимостью обрести максимум общего количества информации;
- пропускная способность дискретного канала $C(X,Y)$ как $max_{P(X)}\big(I(X,Y)\big)$.

Для многих практических задач этап расчета максимума функции $I(X,Y)$ по $P(X)$ является тривиальной задачей, поскольку симметрия элементов матрицы вероятностей перехода зачастую предполагает симметрию входных вероятностей, и, значит, максимизация по нескольким параметрам может быть выполнена на базе численных методов. В то же время алгоритма поиска аналитического решения может быть не столь очевиден и, таким образом данный вопрос является актуальным как в области работы с практическими задачами, так и с точки зрения развития фундаментальной науки.

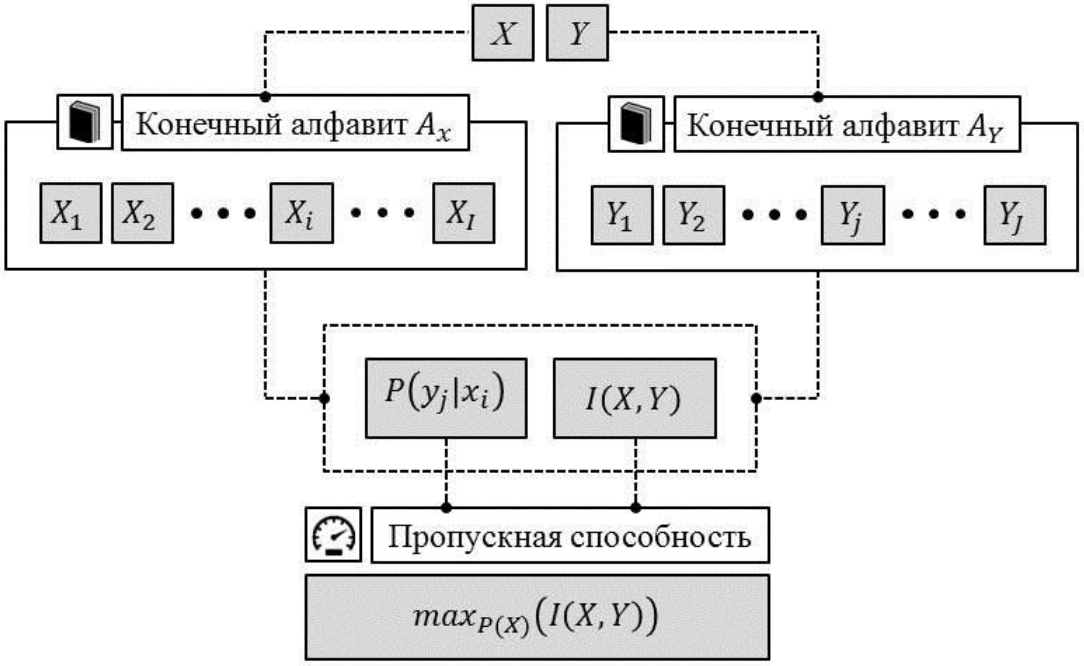

Рис. 1. Схема определения пропускной способности дискретного канала

Рассмотрим случай передачи набора данных через дискретный канал без памяти (рис. 2). Входной набор данных может быть представлен через K символов, которые, соответственно, несут $K \cdot H(X)$ бит информации, причем для их передачи необходимо использовать канал N раз, где N определяется как:

$$N = \frac{K \cdot H(X)}{\mathrm{C}} \qquad (1)$$

Пара (N,K) представляет собой блочный код (block code), т.е. набор векторов длины N, в котором K символов соотносятся с самим информационным блоком, который передается по каналу [3-6]. Таким образом, символы, переданные через дискретный канал без памяти, не являются независимыми. Преобразование данного канала в векторный канал показывает, что каждый вход является одним из векторов передаваемого информационного блока, а значит, вероятность перехода рассчитывается как произведение вероятностей отдельных символов.

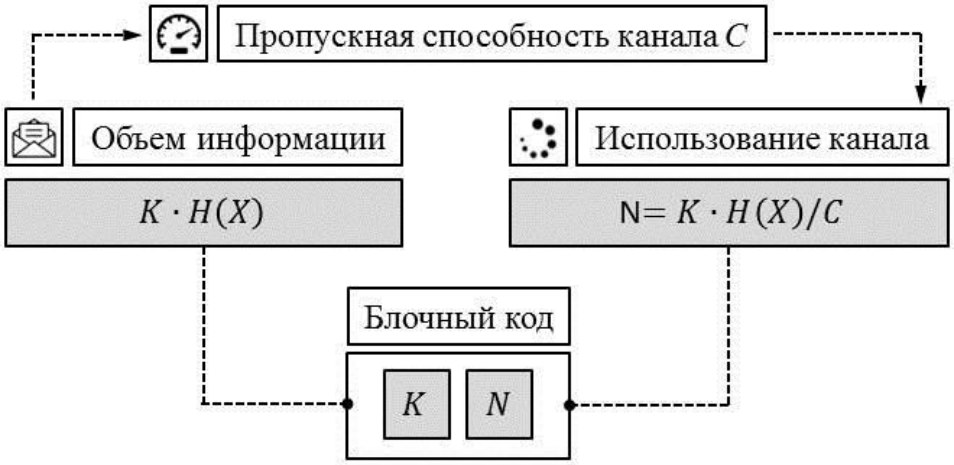

Рис. 2. Схема определения блочного кода для дискретного канала

Если общее количество информации (mutual information) также составляет $k \cdot H(X) = N \cdot C$, а значит $H(X_1, X_2, \dots X_N | Y_1, Y_2, \dots Y_N) = 0$, т.е. выходной вектор должен указывать на уникальность элементов данных в переданном сообщении [1, 7], что может быть рассчитано через одно из уравнений следующей системы:

$$\begin{bmatrix} H(X_1, X_2, \dots X_N) = K \cdot H(X) \\ N \cdot C = H(Y_1, Y_2, \dots Y_N) - N \cdot H(Y|X) \end{bmatrix} \quad (2)$$

Таким образом, выходное распределение будет соответствовать распределению N независимых символов. Вектор, соответствующий передаче данных через модель канала распределяется в набор полученных векторов. В таком случае, целью при построении кода состоит в том минимизации степени перекрытия между наборами. При получении информационного блока происходит этап декодирования и, соответственно, восстановления данных, которые подлежали передачи. В рамках нашей модели рассматривается тот случай, когда все сообщения предполагаются одинаково вероятными. Т.е., вектор y декодируется в вектор x, причем значение $P(y|x)$ должно быть максимальным, что соответствует решающему правилу по методу максимального правдоподобия (maximum-likelihood decision).

2. Методы декодирования линейного кода с ошибками

На уровне построения математической модели декодирования линейного кода с ошибками [6, 8-10] линейный код (N, K) можно рассматривать как линейное векторное пространство размерности K в пространстве N двоичных векторов. Для решения практических задач также следует указать, что двоичный линейный код информационного блока характеризуется длиной N и несет K информационных битов, а кроме того подразумевает $(N - \text{К})$ проверку четности (количество единиц каждого подмножества линейного кода должно быть четным).

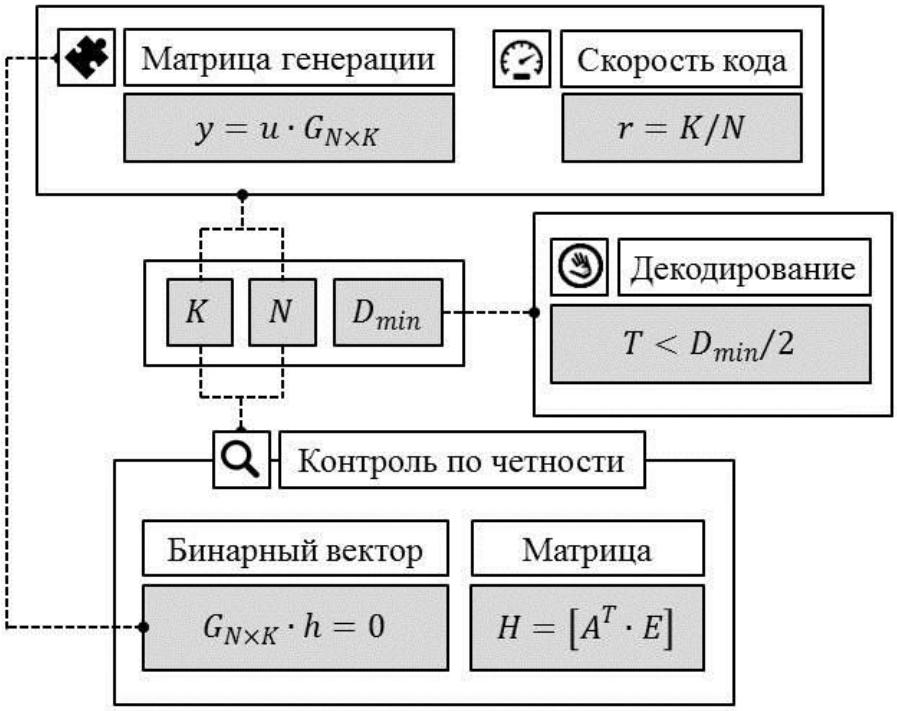

Рис. 3. Математическое моделирование декодирования линейного кода

Для построения методологии работы с линейным кодом необходимо ввести некоторые дополнительные понятия:

- матрица генерации (generator matrix) линейного кода $G_{N\times K}$, как матрица размерности $N \times K$, где линейно независимые векторы k, на основе которых задается линейный код, располагаются в виде строк, а, соответственно, $y = u \cdot G_{N\times K}$;
- скорость линейного кода (rate of a code), которая определяется через соотношение K и N: $r = K/N$;
- единичная матрица $E_{K\times K}$ размерности $K \times K$, через которую определяется матрица генерации как $G_{N\times K} = [E_{K\times K} \cdot A]$;
- бинарный вектор контроля по четности для линейного кода h, множество которых формирует векторное пространство размерности $(N - K)$, причем $G_{N\times K} \cdot h = 0$;
- матрица контроля по четности для линейного кода размерности $(N - K) \times N$, причем $H = [A^T \cdot E_{K\times K}]$;
- минимальное расстояние D_{min} между двумя кодовыми комбинациями (расстояние Хэмминга), через значение которого определяется максимальное допустимое количество ошибок, при котором возможно декодировать код: $T < D_{min}/2$.

Таким образом, полный набор параметров, определяющих код и возможность его декодирования (рис. 3) представляет собой группу (N, K, D_{min}).

В рамках данной модели для определения линейного кода с ошибками, который поддается декодированию, необходимо также ввести понятие границы Хэмминга (Hamming bound) как функции от группы трех переменных N, K и T [11-13], а также условие для максимального значения этой величины, выраженное через следующее неравенство:

$$\begin{cases} B(N,K,T) = \sum_{t=0}^{T} \binom{N}{T} \\ B(N,K,T) \le 2^{N-K} \end{cases} \qquad (3)$$

Далее необходимо определить распределение весов кода $A(w)$ для полного набора бинарных - векторов. Для упрощения задачи анализа конкретного линейного кода также имеет смысл найти выражение для средних значений по наборам кодов, введя ограничение для w:

$$A(w) \sim \binom{N}{w} \rightarrow \begin{cases} \begin{cases} A(w) \approx 2^{(N-K)} \cdot \binom{N}{W} \\ w > 0 \end{cases} \\ \begin{cases} A(w) = 1 \\ w = 0 \end{cases} \end{cases} \quad (4)$$

Для конкретного набора значений (N, K, w), где $w \to 0$,можно получить $A(w) < 1$, но поскольку значение функции $A(w)$ является целым числом, код не может включать в себя комбинацию символов такого малого веса:

$$\binom{N}{D_{min}} \geq 2^{(N-K)}, \quad (5)$$

что определяет наличие кодов с минимальным весом не менее D_{min}.

3. Вероятность ошибки при декодировании блочного кода

Разработанную математическую модель можно специализировать для решения актуальной задачи по расчету вероятности ошибки при декодировании блочного кода, подготовленного для передачи данных через двоичный симметричный канал (BSC: Binary Symmetric Channel) [14-17].

Рассмотрим блочный код, представленный 2^K векторами длины N. Вероятность передачи по BSC определяется через число комбинаций символов $A(w)$ на расстоянии w. Прецедент ошибки возникает в том случае, когда полученная комбинация символов ближе к другой комбинации, отличной от той, что была передана. Для каждой комбинации символов на расстоянии $w > 0$ вероятность того, что она будет верно определена вычисляется как:

$$\begin{cases} \sum_t \binom{w}{t} \cdot p^t \cdot (1-p)^{w-t} \\ t > w/2 \\ w > 0 \end{cases}. \quad (6)$$

Верхнюю оценку вероятности ошибки можно получить взяв сумму вероятностей ошибок:

$$\begin{cases} P(e) < \sum_w \left(\sum_t \left(A(w) \cdot \binom{w}{t} \cdot p^t \cdot (1-p)^{w-t} \right) \right) \\ t > w/2 \\ w > 0 \end{cases}. \quad (7)$$

Данное неравенство можно упростить для характерного случая $t \cong w/2$:

$$\begin{cases} P(e) < \sum_w \left(A(w) \cdot 2^w \cdot (p - p^2)^{w/2} \right) \\ t \cong w/2 \\ w > 0 \end{cases}, \quad (8)$$

что, в свою очередь, таже можно упростить, через введение функции $F = \sqrt{4p \cdot (1-p)}$, как параметр, хаарктеризующий канал передачи блочного кода:

$$\begin{cases} P(e) < \sum_w (A(w) \cdot F^w) \\ F = \sqrt{4p \cdot (1-p)} \\ t \cong \frac{w}{2};\ w > 0 \end{cases}. \quad (9)$$

Разработанный математический аппарат через введение дополнительных условий и аппроксимации можно использовать для решения конкретных задач, например, задачи определения зависимости вероятности ошибки от значения N.

Выводы

В результате проведенного исследования был разработан математический аппарат, который в дальнейшем может быть использован для построения помехоустойчивых систем передачи и приема оцифрованных данных. В частности были предложены:

- обобщенная схема определения пропускной способности дискретного канала;
- обобщенная схема и математический аппарат, которые могут быть использованы для определения блочного кода для дискретного канала;
- математическая модель декодирования линейного кода, который потенциально может содержать в себе ошибки;
- универсальный алгоритм определения вероятности ошибки при декодировании блочного кода при передаче данных двоичном симметричном канале.

Предложенная методология может быть эффективно использована при работе с линейным кодом в системах передачи оцифрованных данных.

Список литературы

1. *Csiszár I. & Körner J.*, 2015. Information theory: Coding theorems for discrete memoryless systems. Cambridge: Cambridge University Press.
2. *Zhong Y., Alajaji F. & Campbell L.L.*, 2007. Error Exponents for Asymmetric Two-User Discrete Memoryless Source-Channel Systems. *2*007 IEEE International Symposium on Information Theory. doi:10.1109/ isit.2007.4557472.
3. *Sungkar M. & Berger T.*, 2018. Discrete Reconstruction Alphabets in Discrete Memoryless Source Rate-Distortion Problems. 2018 IEEE International Symposium on Information Theory (ISIT). doi:10.1109/isit.2018.8437835.
4. *Lei W., Yizhou G., Fucai Z. & Yong W.*, 2018. The Method to Recognize Linear Block Code Based on the Distribution of Code Weight. 2018 13th APCA International Conference on Control and Soft *Computing (CONTROLO).* doi:10.1109/controlo.2018.8439758.
5. *Jadhao M.G.*, 2012. Performance Analysis of Linear Block Code, Convolution code and Concatenated code to Study Their Comparative Effectiveness. IOSR Journal of Electrical and Electronics Engineering, 1(1), 53-61. doi:10.9790/1676-0115361.
6. *Mei T., Zhang C. & Dai Q.*, 2011. Using Linear Block Code and Concatenated Code to Build (k,n) Threshold Scheme. *2011* International Conference on Internet Technology and Applications. doi:10.1109/itap.2011.6006219.
7. *Zeng Q. & Wang J.*, 2017. Information Landscape and Flux, Mutual Information Rate Decomposition and Entropy Production. doi:10.20944/preprints201710.0067.v1.
8. *Wadayama T.*, 2004. An Algorithm for Calculating the Exact Bit Error Probability of a Binary Linear Code Over the Binary Symmetric Channel. *IEEE* Transactions on Information Theory. 50 (2). 331-337. doi:10.1109/tit.2003.822617.
9. *Vu L. & Hayakawa T.*, 2017. An error-correcting code in delay linear network coding. 2017 11th Asian Control Conference *(ASCC).* doi:10.1109/ascc.2017.8287477.
10. *Ilievska N. & Gligoroski D.*, 2015. Simulation of a quasigroup error-detecting linear code. 2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO). doi:10.1109/mipro.2015.7160311.
11. *Kythe D.K. & Kythe P.K.*, 2012. Algebraic and stochastic coding theory. Boca Raton, FL: CRC Press.
12. *Badole V. & Udawat A.*, 2012. Design of Multidirectional Parity Code Using Hamming Code Technique for Error Detection and Correction. Paripex - Indian Journal of Research. 3 (5). 79-81. doi:10.15373/22501991/may2014/27.
13. *Hyun J.Y. & Kim H.K.*, 2004. The poset structures admitting the extended binary Hamming code to be a perfect code. Discrete Mathematics. 288 (1-3), 37-47. doi:10.1016/j.disc.2004.07.010.
14. *Schofield M., Ahmed M.Z., Stengel I. & Tomlinson M.*, 2016. Intentional Erasures and Equivocation on the Binary Symmetric Channel. 2016 International Computer Symposium (ICS). doi:10.1109/ics.2016.0053.
15. *Wacker H.D. & Borcsok J.*, 2007. Probability of Undetected Error on a Binary Symmetric Channel without Memory via Bayesian Inference. Second International Conference on Systems (ICONS07). doi:10.1109/icons.2007.40.
16. *Bao L., Skoglund M. & Johansson K.*, 2006. Encoder¿Decoder Design for Feedback Control over the Binary Symmetric Channel. 2006 IEEE International Symposium on Information Theory. doi:10.1109/isit.2006.262056.
17. *Hagiwara M., Fossorier M.P. & Imai H.*, 2010. LDPC codes with fixed initialization decoding over binary symmetric channel. 2010 IEEE International Symposium on Information Theory. doi:10.1109/isit.2010.5513630.