符号理論

「符号化技術」実験

Dec 2018

1　前書き

符号理論 (Coding Theory) は、情報を符号化して通信を 行う際の効率と信頼性についての理論である。符号は、データ圧縮・暗号 化・誤り訂正・ネットワーキングのために使用される。符号理論は、効率的 で信頼できるデータ伝送方法を設計するために、情報理論・電気工学・数 学・言語学・計算機科学などの様々な分野で研究されている。通常、符号理 論には、冗長性の除去と、送信されたデータの誤りの検出・訂正が含まれ る。

符号化は、以下の 4 種類に分けられる 。

1. 情報源符号化 (source coding) : データ圧縮
2. 通信路符号化 (channel coding) : 誤り検出訂正
3. 暗号符号化 (cryptographic coding)
4. 伝送路符号化 (line coding)

情報源符号化 (データ圧縮) は、データをより効率的に送信するために、情報源からデータを圧縮しようとする。例え ば、ZIP データ圧縮では、データファイルを小さくしてインターネットトラフィックを削減する。データ圧縮と誤り訂正 は、組み合わせて検討することができる。通信路符号化 (誤り検出訂正) は、通信路上に存在する雑音などの障害への耐性を強化するために、余分なデータビッ ト (冗長ビット) を追加する。この技術はあまり目立たないが、例えば音楽 CD ではリード・ソロモン符号を使って傷や 埃による誤りを訂正している。この場合の通信路は CD 自体である。携帯電話も高周波転送におけるノイズや減衰による 誤りを検出訂正する技術を使っている。一般にデジタル信号による通信には、必ず何らかの誤り検出訂正技術が使われ ている。

2　符号理論の歴史

1948 年、クロード・シャノンは論文「通信の数学的理論」を、Bell System Technical Journal の 7 月号と 10 月号の 2 つの記事 で発表した。この論文は、送信者が送信したい情報を最適に符号化する方法の問題に焦点を当てている。この論文で は、ノーバート・ウィーナーが開発した確率論を使用した。当時、確率論は通信理論にはほとんど適用されていなかっ た。シャノンは、メッセージの不確実性の尺度として情報エントロピーを開発し、情報理論の分野を本質的に創始し た。1949 年に二元ゴレイ符号が開発された。これは、24 ビットワードごとに最大 3 つの誤りを訂正し、4 つ目を検出すること ができる誤り訂正符号である。1968 年、リチャード・ハミングは、ベル研究所在籍中の成果である数値計算方法、自動符号化システム、誤り検出訂正 符号でチューリング賞を受賞した。彼は、ハミング符号、ハミング窓、ハミング数、ハミング距離という概念を発明し た。

3　情報源符号化:データ圧縮

情報源符号化の目的は、情報源におけるデータをより小さくすることである。

データ圧縮とは、あるデータをそのデータの実質的な性質（専門用語では「情報量」）を保ったまま、データ量を減らした別のデータに変換すること。高効率符号化ともいう。アナログ技術を用いた通信技術においては通信路の帯域幅を削減する効果を得るための圧縮ということで帯域圧縮ともいわれた。デジタル技術では、情報を元の表現よりも少ないビット数で符号化することを意味する。

データ圧縮には大きく分けて可逆圧縮と非可逆圧縮がある。というより正確には非可逆圧縮はデータ圧縮ではない。可逆圧縮は統計的冗長性を特定・除去することでビット数を削減する。可逆圧縮では情報が失われない。非可逆圧縮は不必要な情報を特定・除去することでビット数を削減する。

しかしここで「不必要な」とは、例えば MP3 オーディオの場合「ヒトの聴覚では通常は識別できない」という意味であり、冒頭の「情報量を保ったまま」という定義を破っている。データファイルのサイズを小さくする処理は一般にデータ圧縮と呼ばれるが、データを記録または転送する前に符号化するという意味では情報源符号化である。

圧縮は、データ転送におけるトラフィックやデータ蓄積に必要な記憶容量の削減といった面で有効である。しかし圧縮されたデータは、利用する前に伸長（解凍）するという追加の処理を必要とする。つまりデータ圧縮は、空間計算量を時間計算量に変換することに他ならない。例えば映像の圧縮においては、それをスムースに再生するために高速に伸長（解凍）する高価なハードウェアが必要となるかもしれないが、圧縮しなければ大容量の記憶装置を

必要とするかもしれない。データ圧縮方式の設計には様々な要因のトレードオフがからんでおり、圧縮率をどうするか、(非可逆圧縮の場合) 歪みをどの程度許容するか、データの圧縮伸長に必要とされる計算リソースの量などを考慮する。

新たな代替技法として、圧縮センシングの原理を使ったリソース効率のよい技法が登場している。圧縮センシング技法は注意深くサンプリングすることでデータ圧縮の必要性を避けることができる。

4 通信路符号化:誤り検出訂正

通信路符号化の目的は、なるべく高速に転送でき、なるべく多くの符号語を含み、誤り検出訂正可能な符号を見出すことである。これらの目的は互いに相反するため、用途によって適切な符号体系は異なる。符号に求められる特性は、転送中に発生するエラーの確率に依存する。

例えば、CD では埃や傷による誤りを訂正することを主に考慮している。従って符号はインターリーブされた形式となり、データはディスク面のあちこちに分散される。よい符号とは言えないが、単純な繰り返し符号を例として考える。例えば、何らかの (音声のような) データのブロックを 3 回送信するとする。受信側は 3 回受信したデータブロックをビット毎に比較し、多数決で正しいデータを決定する。これを少しひねって、ビットの送信順を変えてインターリーブさせる。データを 4 つの小さいブロックに分割し、1 つめのブロックの 1 ビット目の次に 2 つめのブロックの 1 ビット目という順に送信するのである。これをディスク面全体に分散するよう 3 回繰り返す。このような単純な繰り返し符号ではあまり効率的ではないが、実際にはもっと効率的な符号を使って情報をインターリーブし、ディスク面の一部に傷があっても誤り訂正できるようにしている。

別の用途にはもっと適した符号が別に存在する。宇宙空間での通信は受信機の熱雑音の影響が大きく、これは CD の傷などとは異なり、連続的なノイズである。電話回線を使ったモデムではノイズがあるために転送速度が制限されるが、それと同様である。携帯電話は減衰が問題となる。高周波では受信機がほんの数センチ動いただけでも減衰により信号が捕らえられなくなる。このような減衰に対処する通信路符号化の技法も存在する。

代数的符号理論 (Algebraic coding theory) とは、符号の特性を代数学的に表現し研究する分野である。代数的符号理論は基本的に以下の 2 つの符号に分類される。

- 線型ブロック符号
- 畳み込み符号

主に符号の以下の特性を分析する。

- 符号語の長さ
- 正しい符号語の総数
- 2 つの正しい符号語間の最小ハミング距離

4.1 線型ブロック符号

線型ブロック符号は線型性を有している。すなわち、任意の 2 つの符号語の総和も符号語であり、情報源のビット列のブロックにもそれが適用される（そのため線型ブロック符号と呼ぶ)。線型でないブロック符号も存在するが、それでよいかどうかを証明することは困難である。線型ブロック符号は (n,m,d_{min}) で表され、それぞれ以下のような意味を持つ。

- n は符号語の長さ（シンボル数）
- m は一度に符号化されるシンボル数
- d_{min} は符号間の最小ハミング距離

線型ブロック符号に属する符号として以下のようなものがある。

- 巡回符号（ハミング符号は巡回符号のサブセット）
- 反復符号
- パリティ符号
- リード・ソロモン符号
- BCH 符号
- 代数幾何符号
- リード・マラー符号
- 完全符号

ブロック符号は、硬貨を敷き詰める問題と関係している。これは 2 次元で考えると分かりやすい。硬貨を何枚もテーブルの上に並べ、なるべく稠密に敷き詰める。すると、ちょうど蜂の巣のように正六角形状に敷き詰められる。しかし、ブロック符号はもっと高次元であり、容易に視覚化できない。宇宙空間での通信に使われた強力なゴレイ符号では 24 次元を使っている。一般的な 2 進数の符号では次元は符号語の長さとなる。符号理論では、N 次元球モデルを使う。例えば、テーブル上の円に何枚の硬貨を敷き詰められるか、あ

るいは 3 次元では球体の中にどれだけビー玉を詰められるかという問題と同じである。別の考慮として、符号の選択がある。例えば、正六角形を四角形の枠に敷き詰めようとしても、角に隙間ができてしまう。次元を大きくすると、隙間となる空間の割合は小さくなる。しかし、ある次元で符号が隙間無く敷き詰められるようになり、それを完全符号と呼ぶ。そのような符号の例は非常にまれである（ハミング [n,k,3]、ゴレイ [24,12,8],[23,12,7],[12,6,6]）。

4.2 畳み込み符号

畳み込み符号は電話回線用モデム (ITU-T V.32、V.17、V.34) や GSM 携帯電話、さらには衛星通信や軍事通信機器にも使われている。

ここでのアイデアは、入力となるメッセージ群のシンボル列の重み付き総和として各符号語のシンボルを作成するということである。これは線形時不変系において入力とインパルス応答が判っているときに出力を求める畳み込みに似ている。

従って、畳み込みエンコーダの出力は一般に、畳み込みエンコーダとレジスタの状態に対する入力ビットの畳み込みである。基本的に畳み込み符号は同等なブロック符号以上のノイズ耐性を保証しないが、多くの場合、同程度のブロック符号よりも実装が大幅に単純化される。エンコーダは大抵の場合、状態メモリとフィードバック論理（通常 XOR ゲート）を持つ単純な回路である。デコーダはファームウェアやソフトウェアで実装される。畳み込み符号のデコードに最適なアルゴリズムとしてビタービ・アルゴリズムがある。その計算負荷を減らす単純化手法もあり、最も可能性の高い経路だけを探索する。これは最適ではないが、低ノイズの環境ではよい結果となることがわかっている。最近のマイクロプロセッサでは、この縮小された探索アルゴリズムで平均毎秒 4,000 符号語以上のデコードが可能である。

5 暗号符号化

暗号および暗号符号化は、第三者（敵対者（英語版））の存在下で安全な通信を行うための技術である。より一般的には、敵対者をブロックする通信プロトコルの構築と分析に関するものである。データの機密性と完全性、認証、否認防止などの情報セキュリティのさまざまな側面が、現代の暗号の中心である。現代の暗号は、数学、コンピュータ科学、電気工学の分野の境界上に存在する。暗号化を応用したものには、ATM カード、コンピュータパスワード、電子商取引などがある。

6　伝送路符号化

伝送路符号（デジタルベースバンド変調またはデジタルベースバンド送信方法とも呼ばれる）は、データ伝送路を介してデジタル信号を伝送する際に、デジタル信号をデータ伝送路の特性に適した電圧・電流または光子のパルス波形に変換するための符号である [5]。伝送路符号は、デジタルデータ転送によく使用される。伝送路符号は、デジタルデータ転送によく使用される。伝送路符号は、搬送されるデジタル信号を、物理チャネルおよび受信装置の特性に応じて最適に調整された振幅および時間離散信号によって表すことからなる。デジタルデータの 1 と 0 を表すために使用される電圧または電流の波形パターンを、伝送路符号という。伝送路符号の一般的なタイプは、単極符号（英語版）・両極符号（英語版）・マンチェスタ符号である。

7　符号理論の他の応用

符号理論におけるもう 1 つの課題は、同期を可能とする符号の設計である。例えば、位相変移 (phase shift) を容易に検出・訂正できるよう符号を設計すれば、複数の信号を同じ通信路で同時に送ることができる。例えば、携帯電話で使われている符号分割多元接続 (CDMA) 符号がある。その詳細は本項目の範囲外だが、大まかに言えば、各携帯電話に特別な符号語が割り当てられる。転送時、その符号語を使って音声を表すビット列をスクランブル（暗号化）する。受信機では、その逆を行って暗号解読する。このような符号語の特性により、同時に複数の携帯電話がそれぞれ個別の符号語を割り当てられ、通話可能となる。1 つの受信機から見れば、他の通話の信号は低レベルなノイズとしか認識されない。

もう 1 つの一般的な符号のクラスとして、自動再送制御 (ARQ) 符号がある。この場合、送信機は長いメッセージにパリティビットを付与する。受信機はメッセージとパリティビットが一致するかを調べ、一致しない場合に送信機にメッセージの再送を要求する。ごく単純なものを除いて、Wide Area Network で使用されるプロトコルには必ず ARQ が使われている。例えば、SDLC (IBM)、TCP、X.25 などである。この分野では、拒絶されたパケットと新たなパケットの一致問題という部分でも研究が進んでいる。

つまり、新たに受信したパケットが再送されたものか、それとも別の新しいパケットなのかを識別する問題である。一般にパケットに番号を振ることで対処するが、プロトコルスタックがある場合、再送を制御する階層が異なる場合がある。TCP/IP は両方の技法を採用している好例である。コネクションのある場合、TCP/IP は ARQ 符号による再送を行う。しかし、コネクションがない場合、ARQ は使われず、アプリケーション層で（必要に応じて）再送を制御しなければならない。