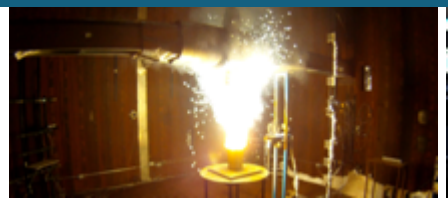
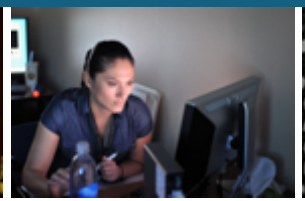
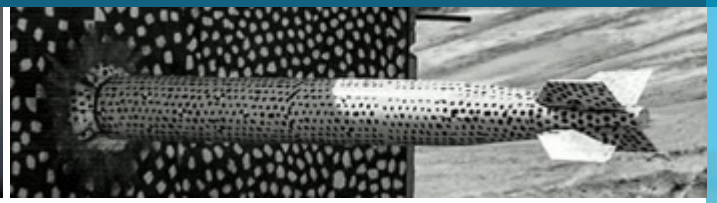
SAND2022-9696C

Sandia National Laboratories

What is the Best Model? A Statistical Approach to Compression Analytics







CIS LDRD 225949 (Year 1 of 2)
Alex Foss (PI, 5573), Justin Newcomer (PM, 5570)

Sandia Team: Christina Ting (5554), Kurtis Shuler (5573), Rich Field (5553), Travis Bauer (5554)
UIUC AA: Eddie Cardenas-Torres, Dave Zhao

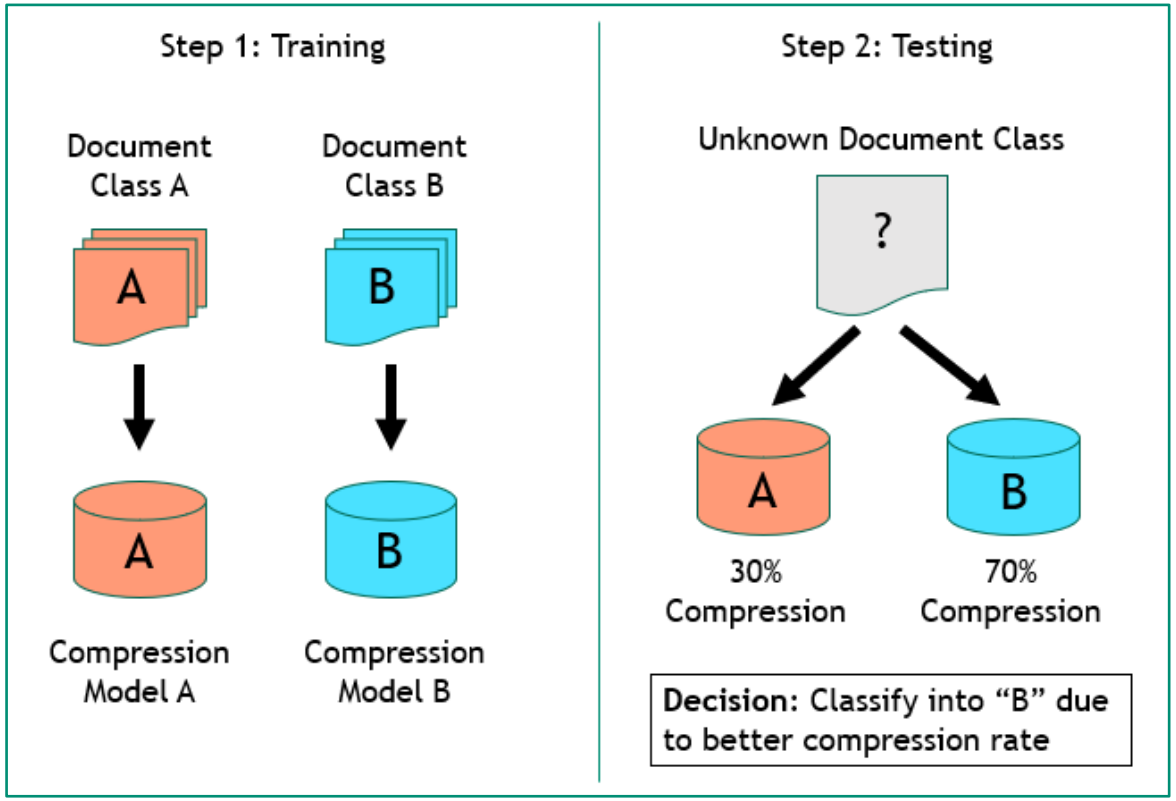
U.S. DEPARTMENT OF ENERGY NNSA

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia LLC, a wholly owned subsidiary of Honeywell International Inc. for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

PROJECT BACKGROUND

- **Definition:** Compression *analytics* (CA) is the use of compression *algorithms* for machine learning (ML)
  - Wide range of applications: Text, raw binaries, seismic measurements, genomics, cyber log data
- **Problem:**
  1. Lack of formal statistical models for CA limit principled extensions
  2. Unclear how to select best CA model
  3. Lack of rigorous ways to optimize misclassification rate
- **Goals:**
  1. Develop novel statistical framework for CA
     - Statistical model selection
     - Optimal classification rules
  2. Use framework to create novel CA techniques
     a) Develop nonlocal CA



Goal: Develop novel statistical and nonlocal CA algorithms for ML

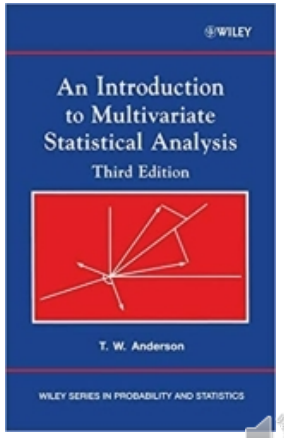
Background: Minimizing the Expected Cost of Misclassification

Theorem for deriving theoretically optimal classification rules:

Theorem: For a K-class classification problem, the regions $R_1, R_2, \ldots, R_K$ that minimize the expected cost of misclassification are given by

$$R_k = \left\{ x : \sum_{i=1}^{K} \pi_i P(x \mid \text{class } i) cost(k \mid i) < \sum_{\ell=1}^{K} \pi_\ell P(x \mid \text{class } \ell) cost(j \mid \ell) \right\} \quad \forall j = 1, 2, \ldots, K, \quad j \neq k$$

- x is a scalar or vector of observed covariates
- $P(x|\text{ class} = i)$ is the PMF/PDF of the random variable X evaluated at x within class i
- π_i is the prior probability of observing class i
- $cost(j|i)$ gives the cost of (mis)classifying into class j given true class i
- Can be adapted for minimax classification, maximizing accuracy/F_1/etc.



T.W. Anderson. *An Introduction to Multivariate Statistical Analysis, 3rd Ed*, Hoboken, NJ: Wiley.

Result 1: Minimizing the Expected Cost of Misclassification with CA

Consider that

$$-\log_2 P_{\vec{X}}(\vec{x}) = I_{\vec{X}}(\vec{x}), \quad \text{and}$$
$$I_{\vec{X}}(\vec{x}) \approx C(\vec{x} \mid \mathbb{X})$$

where

- $I_{\vec{X}}(\vec{x})$ denotes the information content in the observation $\vec{X} = \vec{x}$
- $C(\vec{x} \mid \mathbb{X})$ denotes the compressed length of x given by a compression model C trained on $\mathbb{X}$, a sufficiently large sample from the random variable X

This allows the approximation of optimal classification rules as

$$R_k \approx \left\{ x : \sum_{i=1}^{K} \pi_i 2^{C(x|\mathbb{X}_i)} cost(k \mid i) < \sum_{\ell=1}^{K} \pi_\ell 2^{C(x|\mathbb{X}_\ell)} cost(j \mid \ell) \right\} \quad \forall j = 1, 2, ..., K, \quad j \neq k,$$

where the approximation error is determined by the approximation of $-\log_2 P_{\vec{X}}(\vec{x})$ with $C(\vec{x} \mid \mathbb{X})$.

Novel application of optimal classification rules to compression analytics classifiers.

Background: Prediction by Partial Matching (PPM) and Model Order

PPM is a commonly used compression algorithm (e.g. RAR, 7-zip)

$$C_{PPM(K)}(\vec{x}\,|\mathbb{X}) \approx \prod_{i=1}^{n} \mathrm{P}(x_i \mid x_{i-K}, x_{i-K+1}, \ldots, x_{i-1})$$
$$\approx \mathrm{P}(x_1, x_2, \ldots, x_n)$$

- Complete PPM algorithm not shown here (See Cleary & Witten and Bauer 2021)
- $\vec{x}$ is a vector of observed symbols drawn from a vector of random variables $\vec{X}$
- $\mathbb{X}$ is a sufficiently large sample drawn from $\vec{X}$ used to estimate the required probabilities
- Main idea: PPM attempts to estimate a **variable order markov model** (VMM)
  - Requires choosing a maximum context length (order) K
  - Includes a recursive method for defaulting to context length $K-1$ for an unobserved context K
- Many PPM variants exist

Cleary, J.G. and Witten, I.H. (1984) Data Compression Using Adaptive Coding and Partial String Matching. IEEE Transactions on Communications, 32, 396-402.
Bauer, T. (2021) Ngram PPM: Compression Analytics without Compression. Sandia Report, SAND2021-11185.

Background: Model Selection with AIC/BIC

A statistical model $P_X(x\,;\theta)$ with parameters θ has likelihood

$$L(\theta;\vec{x}) = \prod_{i=1}^{n} P_X(x_i;\theta)$$

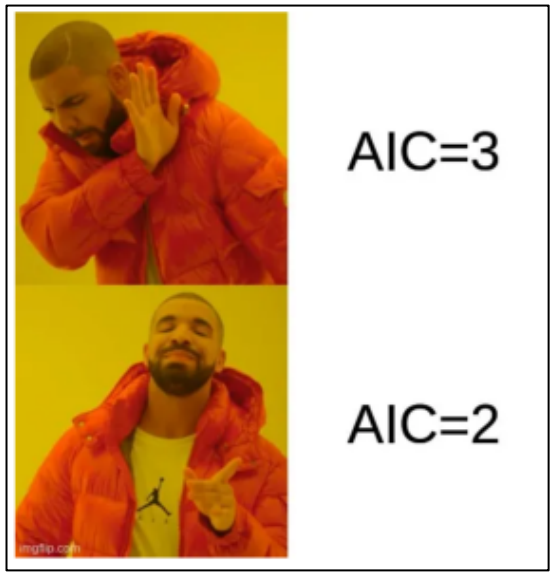
We can evaluate model quality via AIC (Akaike's Information Criterion):

$$AIC(\hat{\theta}) = 2|\hat{\theta}| - 2\ln(L(\hat{\theta};\vec{x}))$$

$$BIC(\hat{\theta}) = \ln(n)|\hat{\theta}| - 2\ln(L(\hat{\theta};\vec{x}))$$

$$\downarrow AIC/BIC \;\Rightarrow\; \uparrow \text{Model Quality}$$

where $\hat{\theta}$ is an estimator of θ, and $|\hat{\theta}|$ is the dimensionality of $\hat{\theta}$.



AIC/BIC are principled methods for selecting parsimonious statistical models.

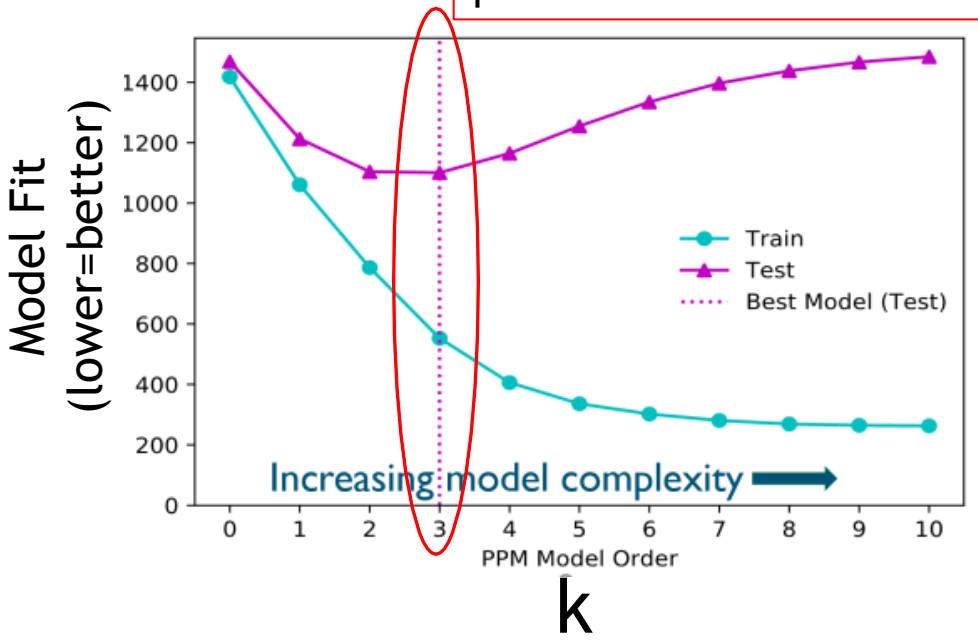
Result 2: Choosing PPM Model Order with AIC/BIC

For PPM encoding, AIC/BIC can be used to select model order k

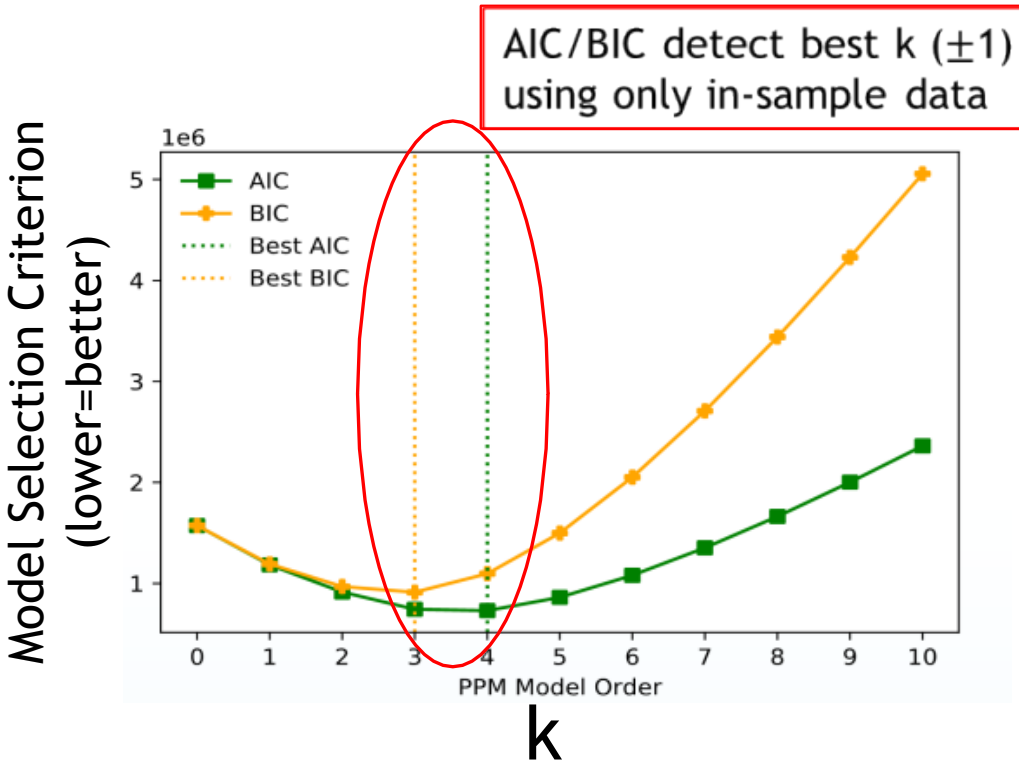
- $|\theta|$ can be calculated as the number of estimated probabilities
- Likelihood can be approximated using the compressed length of x

Sample results on 20-Newsgroups data:

k=3: Best observed out-of-sample performance



AIC/BIC detect best k (±1) using only in-sample data



Novel AIC/BIC implementation gives principled technique for model selection with PPM

Conclusions

- Compression Analytics (CA) involves using compression algorithms for machine learning (ML)
- Statistical interpretation of CA offers a path forward for novel method development
- Result 1: Optimal classification rules adapted to general compression algorithms
- Result 2: Statistical model selection for choosing context length in PPM
- Future directions:
  - Nonlocal compression for extending context length (order)
  - Adapting latent variable methods for unsupervised CA

Thanks for listening!