Structure validation in clustering by stability analysis

Joachim M. Buhmann

Computer Science Department, ETH Zurich

Learning discrete structures under uncertainty

- **Partitions or clusterings**: Which model? How many clusters?
- **Trees or dendrograms**: What depth of the tree? How many leaves?
- **Graphical models**: How many edges?
Do you recognize the object?

What is information?

Is Shannon information useful as the technical measure of information?
Relevant information in image – where can we find it?

The relevance of information is determined by the task!

Yarbus (1967)

Free examination
1

Estimate material circumstances of the family
2

Give the ages of the people.
3

Surmise what the family had been doing before the arrival of the unexpected visitor.
4

Remember the clothes worn by the people.
5

Machine learning & statistical modelling

- **Approximate optimization**, stability and information theory
  - optimization as coding
  - statistical models

- **Medical image processing** of tissue data for high throughput cancer diagnosis
  - **computational pathology**: Tissue Microarrays for clear cell renal carcinoma
What is the „right“ model for figure ground segmentation?

Path Based Clustering

Clustering models for figure ground segmentation?

**Compactness** criterion
- K-Means Clustering
- Pairwise Clustering, AvAssoc
- Correlation Clustering
- Max-Cut, Average Cut
- Normalized Cut

**Connectedness** criterion
- Single Linkage
- Path Based Clustering
Design problems in clustering: validation

- **Modeling problem**: Does the cluster model “describe” the data? Selection of the costs/hypothesis class!

- **Model order selection problem**: Is the number of clusters and/or features correct?

Concept for robust data analysis

- *Data vectors, relations, images,...*
- *Structure definition (costs, risk, ...)*
- *Structure optimization multiscale analysis, stochastic approximation*
- *Structure Validation statistical learning theory*
- *Regularization of statistical & computational complexity*
- *Quantization of solution space Information/Rate Distortion Theory*
- *Feedback*
Mathematical formalization of clustering

- Given: **object space** $\mathcal{O}$ with **objects** $o \in \mathcal{O}$.
- Given: **measurement space** $\mathcal{X}$
- **data** are relations $(o, X) \in \mathcal{O} \times \mathcal{X}$
- Clusterings **partition** objects into groups, i.e.,

$$c : \mathcal{O} \times \mathcal{X} \rightarrow \{1, \ldots, k\}$$

$$(o, X) \mapsto c(o, X)$$

- **Hypothesis class** $c \in \mathcal{C} \equiv \{\text{partitions of data}\}$

Order relation of clusterings

- **algorithm** $\alpha$ selects “statistically optimal” clusterings $\alpha : \mathcal{O} \times \mathcal{X} \rightarrow \mathcal{C}_\gamma \subset \mathcal{C}$

- **Remark**: $\alpha$ could minimize a cost function where $\mathcal{C}_\gamma$ is a $\gamma$ close approximation of the minimum.

- **Model selection problem**: Which properties should a good clustering algorithm $\alpha$ possess?

- **Stability!** Small changes of data should yield similar clusterings.
Why risk approximation?

- **Data often contain noise!** Very frequently **data** are best modeled as **random variables**.
- **An empirically optimal** clustering often is statistically indistinguishable from other **equally plausible** data partitionings!
- **Data noise reduces resolution in data space!**

=> This quantization induces a **quantization of the hypothesis class** for structures.

Empirical Risk Approximation

- **Learning**: sample typical solutions of an approximation set \( C^{(1)}_\gamma \equiv C_\gamma (X^{(1)}) \) given data \( X^{(1)} \)

\[
C^{(1)}_\gamma \equiv \{ c : d (c(X^{(1)}), c(X^{(1)})) \leq \gamma \}
\]

- **Algorithm**: Gibbs sampling of clusterings with temperature \( T(\gamma) \) explores approximation set.
- **Interpretation**: \( T(\gamma) \) controls the resolution of the hypothesis class, i.e., the **minimal similarity** of statistically indistinguishable structures.
Approximate optimization and information theory

- **Problem**: Data processing in Machine Learning is often formulated as an optimization question.

  => noisy data require **robustness = generalization**

- Use **approximate optimization results** as code

- “Communication” is achieved via **approximate optimization of instances** since test instances are considered to be perturbed training instances.

---

**Two instance setting for learning**

- **Instance space (data space)**
  - $2^n$ states for graph cut

- **Solution spaces (Hypothesis class)**
  - $2^n$ training solutions

- **Instance/data space**
  - $2^n$ states

- $x$: ERM solutions

- $\phi(c)$
**Figure Ground Segmentation by Graph Cut**

Graph cut code problems

Graph cut test problem

**Robust Approximations by Stability**

Define a set of code problems

Problem generator PG

$R(\cdot, X^{(1)})$

$R(\cdot, X^{(1)})$

$\{\sigma_1, \ldots, \sigma_{2^M}\}$

Sender

Receiver

$X^{(1)}$

$c^\perp = \arg \min_c R(c, X^{(1)})$

$\sigma \circ X^{(1)}$

$c^\perp (\sigma \circ X^{(1)})$
Robust Approximations by Stability

1. Sender sends a permutation index $\sigma_s$ to problem generator.
2. Problem generator sends a new problem with permuted indices to receiver without revealing $\sigma_s$.
3. Receiver identifies the permutation $\sigma^*$ by comparing approximation sets.

Communication Process

- Receiver has to compare sets of clusterings $C_\gamma(X^{(1)})$ of training instance (code problem) with approximate clusterings $C_\gamma(X^{(2)})$ of the test data.
- Define a mapping $\phi : C(X^{(2)}) \rightarrow C(X^{(1)})$
- Decoding

\[
\sigma^* = \arg\max_{\sigma} \left| C_\gamma(\sigma \circ X^{(1)}) \cap \phi \left( C_\gamma(\tilde{X}^{(2)}) \right) \right|
\]

if
\[
\frac{|C_\gamma(\sigma^* \circ X^{(1)}) \cap \phi \left( C_\gamma(\tilde{X}^{(2)}) \right)|}{|C_\gamma(\sigma \circ X^{(1)})|} \geq 1 - \epsilon
\]
Error Analysis and Approximation Capacity

- Approximations of sender and receiver have little in common! => **Irreproducibility**
  This condition determines approximation precision

- **Approximations** of test problem has a large overlap with approximations of wrong training problem! => **Confusion**

=> Select model with maximal information capacity, i.e., high precision and high noise robustness

---

Approximation Capacity

- **Condition of vanishing total error**

\[
M \log 2 < \log |\phi \circ C_\gamma(\sigma_s \circ X^{(2)}) \cap C_\gamma(\sigma_j \circ X^{(1)})| \\
- \log |C_\gamma(\sigma_s \circ X^{(2)})| - H(\sigma_s) \\
≡ I(\sigma_s, C_\gamma(\sigma_s \circ X^{(2)}))
\]

- **Model selection**: Maximize the mutual information w.r.t. **topology** of solution space, **metric** of solution space, **cost** function, transfer function \(\phi\) and approximation **precision** \(\gamma\).
Results on Toy Data

Clustering of Microarray Data

(dataset from Golub et al., Science, Oct. 1999, pp.531-537)

**Task:** Find groups of different Leukemia tumour samples (two- and three class classifications are known).

**Problem:** Number of groups is unknown a priori.

**Via Stability with k-means:** Estimated number of groups is 3.

**Result:** 3-means solution recovers 91% of known sample classifications.

3-means grouping of Golub et al. data set and estimated instability

Via Stability with \(k\)-means: Estimated number of groups is 3.
Scales in Data Analysis and Vision

- Refinement of Variable Space
- Refinement of Optimization Criterion
- Refinement of Model Order

- Increment Level of Resolution Pyramid
- Increase Regularization
- Increase # of Segments

Machine learning & statistical modelling

- Approximate optimization, stability and information theory
  - optimization as coding
  - statistical models

- Medical image processing of tissue data for high throughput cancer diagnosis (MICCAI '08)
  - computational pathology: Tissue Microarrays for clear cell renal carcinoma
Renal (Clear) Cell Carcinoma

Renal cell carcinoma (RCC) is one of the ten most frequent malignancies in Western societies.

The prognosis of renal cancer is poor.

Many patients suffer already from metastases at first diagnosis.

The identification of biomarkers for prediction of prognosis (prognostic marker) or response to therapy (predictive marker) is therefore of utmost importance to improve patient prognosis.

Tissue Microarray Preparation

Hematoxylin-Eosin Staining

Immunohistochemistry
Tissue Microarray preparation

- Case #1
- Paraffin recipient

Tissue Microarray Analysis

- Stained TMA slides
- Immunohistochemistry identifies potential biomarker (e.g. MIB-1)

Survival (%)
- 100
- 90
- 80
- 70
- 60
- 50
- 40
- 30
- 20
- 10
- 0

- p<0.0001

Months
- 100
- 80
- 60
- 40
- 20
- 0

Expressed
- not expressed
Data Analysis Problem: Variability

H & E

MIB-1

Tissue Microarray Analysis

Nuclei Detection  Nuclei Classification

5%

Nuclei Segmentation  Staining Estimation

<table>
<thead>
<tr>
<th>Spot</th>
<th>Staining</th>
<th>Grading</th>
</tr>
</thead>
<tbody>
<tr>
<td>15</td>
<td>4%</td>
<td>G1</td>
</tr>
<tr>
<td>16</td>
<td>35%</td>
<td>G1</td>
</tr>
<tr>
<td>17</td>
<td>9.5%</td>
<td>G3</td>
</tr>
</tbody>
</table>
TMA Annotator

Gold-Standard by expert classifications

- 2 pathologists
- ~2500 nuclei
- 15% detection mismatch
TMA Classifier – Tumor / Normal Labels

Agreement among pathologists

180 cell nuclei are randomly selected out of 9 areas.

agreement for 105 cell nuclei.

number of cell nuclei

no cancer cancer

3 2 1 -1 -2 -3

consistency among pathologists self confusion matrix of pathologists
Workflow of computational pathology

1. training examples
2. feature extraction
3. Random Forest learning
4. cell nuclei detection
5. estimate staining
6. application to patient cohort
7. estimated Marker distribution
8. survival analysis

Prediction of survival

- The learning algorithm estimates the number of stained cancer cell nuclei more reliably than a trained pathologist for patients with good prognosis.

- Kaplan-Meier curve for:
  - 33% low risk patients
  - 33% intermediate risk
  - 33% high risk patients

- reliability of diagnosis will be increased!
Improving the quality of training data

- **Problem**: Labeling information by pathologists is often inconsistent! **No ground truth**
- **Solution**: Filter out samples which are hard to classify, i.e., denoising
- **Strategy**
  1. Compute how similar the samples appear to the trees of the random forest.
  2. Cluster the samples into groups of high similarity.
  3. Analyze label inconsistency in these groups.

Wishart-Dirichlet Cluster Process …

- … is a sequence of inner product matrices of growing size, and a random partition $B$ of objects into $k$ blocks.
- **Dirichlet-Multinomial prior** over partitions
  \[
  P_n(B | \lambda, k) = \frac{k!}{(k-k_B)!} \frac{\Gamma(\lambda) \prod_{b \in B} \Gamma(n_b+\lambda/k)}{\Gamma(n+\lambda)[\Gamma(\lambda/k)]^{k_B}}.
  \]
- Similarity matrices are Wishart distributed
  \[
  S|B \sim \mathcal{W}_d(\Sigma_B) \quad \text{with} \quad \Sigma_B = I_n \otimes \Sigma_0 + B \otimes \Sigma_1.
  \]
Application: Automated TMA Analysis

Dataset: 500 cancerous and 500 normal nuclei from TMA spots of renal clear cell carcinoma patients.

Malignant/benign classification by Random Forest yields 36% error.

Enhancement: search for subgroups of highly discriminative nuclei. Similarity matrix is defined by proximity of nuclei in tree ensemble.

Some negative eigenvalues of similarities require to use a shift invariance method.

Classification Results

- Random Forest trained on subset of nuclei from cluster 2 and 3: **19.4% test error** => Importance of quality assessment prior to classification!
- This strategy overcomes a severe problems in the design of computational TMA analysis tools.
Future challenges in Computer Science

- **Machine Learning** has matured to an essential method of computer science, to generate complex statistical models for **Computational Biology**, **Visual Computing** but also other core areas of informatics.

- **Statistical modeling** and **algorithmics** as core problems of computer science can be excellently studied in the area of **Pattern Recognition**.

- We will provide society with the **intelligent technology of the 21st century!**