pertinence
simple solutions for complex processes

www.pertinence.com

Clustering and Prediction: some thoughts

Olivier Bousquet
olivier.bousquet@pertinence.com

Tübingen, August 15th, 2005

> Goal of this talk

- **Not a presentation of research results**
- **But some ideas and questions, in order to stimulate discussions**

Outline

- **Can we build a theory of clustering?**
- **What is clustering?**
- **What is the right model for analyzing it?**
- **What are the right questions?**
- **Conclusion**

pertinence

Towards a theory of unsupervised learning?

- **For supervised learning, statistical learning theory relatively mature**
- **Could similar results be obtained for unsupervised learning?**

pertinence

> What can we hope from such a theory?

- **"It should tell us which are the good algorithms"**
- **"It should allow us to build better algorithms"**
- **"It should help us choose the number of clusters"**
- **Is all this reasonable ???**

pertinence
simple solutions for complex processes

> The negative answer

- **Supervised learning theory does not answer these questions**
  - No free lunch
  - Finite sample size: error can be arbitrarily close to chance
  - Slow rates: no universal rate of convergence
  - Model selection/SRM: using bounds is just a way to incorporate a prior
- **SLT can only**
  - give separate answers to the questions of computation, estimation, approximation
  - but not tell you how to trade them
  - in particular, it cannot tell whether an algorithm is better than another
- **So for unsupervised learning which is not even properly defined, there is not much hope**

pertinence
simple solutions for complex processes

So what should we do?

- **Define the goal**
  - What is clustering, why do we want to do it?
  - Need definitions, principles, axioms
- **Choose the right model**
  - Need a framework for analyzing the algorithms and making statements
- **Ask the right questions**
  - Which questions can possibly be answered within this framework?
  - Answers will come in the form of theorems

pertinence
simple solutions for complex processes

Outline

- **Can we build a theory of clustering?**
- **What is clustering?**
  - What is clustering?
  - What is the quality of a clustering?
  - What is the quality of an algorithm for a problem?
  - What is the quality of an algorithm in general?
- **What is the right model for analyzing it?**
- **What are the right questions?**
- **Conclusion**

pertinence
simple solutions for complex processes

> What is the task of clustering?

- **"Extract hidden structure in the data"**
- **"Get a simple yet meaningful description of the distribution of the data" (S. Ben-David)**
- **"At which scale does the music play in the data?" (J. Buhmann)**

- **Why do we need clustering?**
  - **Understanding the data, finding groups**
  - **Feature extraction for classification**
  - **Summarization / Compression**

pertinence
simple solutions for complex processes

> What is a clustering?

Consider a space X (finite or infinite)

- **Clustering = Partition**

$$f : \mathrm{X} \to \{1,\ldots,K\}$$

- **Soft partition: map to the K-simplex**

$$f(x) = (p_1,\ldots,p_K) \text{ with } \sum_{k=1}^{K} p_k = 1$$

- **Hierarchical model: a collection of (nested) partitions for each K in N**

pertinence
simple solutions for complex processes

What should be measured?

- **How good is this partition of the data?**
  - Empirical quality (typically the criterion optimized by the algorithm)
  - Expected quality on future data (requires extension and knowledge about the distribution, or can be estimated by CV or bounds)
- **How well does my algorithm extract structure in this problem?**
  - Several instances can be created from a single problem (sampling)
  - For a given distribution P, estimate how "good" is the structure extracted from P by the algorithm on average
  - Not for a specific clustering
  - Can be used for model order selection
- **How well does my algorithm work?**
  - Assess quality on a class of problems
  - Cannot be used for model order selection

pertinence
simple solutions for complex processes

Outline

- **Can we build a theory of clustering?**
- **What is clustering?**
  - What is clustering?
  - What is the quality of a clustering?
  - What is the quality of an algorithm for a problem?
  - What is the quality of an algorithm in general?
- **What is the right model for analyzing it?**
- **What are the right questions?**
- **Conclusion**

pertinence
simple solutions for complex processes

What is a good clustering?

- **Consider**
  - X random variable representing the data
  - Y random variable distributed according to the (soft)-partition function

$$P(Y = k | X) = p_k(X)$$

- **Y should be such that it retains the (relevant) information/structure contained in X**
- **Structure is related to prediction: X and Y are related if one can be predicted from the other**
- **Need assumptions**
  - Depends on a given loss function
  - Depends on a given prediction algorithm

pertinence

Application to clustering quality

- **First option: from Y to X**
  - "How much information did we loose replacing X by the labels Y?"
  - Measured by the ability of recovering X from Y
  - Deterministic predictor, loss given by d(X,X'), h maps each cluster to a centroid

$$E(d(X, h_n(Y)))$$

  - Typically the kind of measure used in centroid-based clustering
- **Second option: from X to Y**
  - For example, compute the CV error of a predictor on the labeled sample
  - The algorithm that is used encodes the regularity assumptions

pertinence

Second-order structure

- **First-order structure: how X can help predict Y, when given a labeled sample**
  - How smooth is Y with respect to X?
  - How extrapolated value match true values?
- **Second-order structure: how X can help predict Y, when given an unlabeled sample**
  - How smooth is Y with respect to P(X)?
  - How one can extrapolate based on the distribution of X only?
- **Combining both**
  - Need a semi-supervised algorithm
  - Take average error given n labeled and m unlabeled points

Extension

- **The quality can be measured on points outside the dataset**
  - Some algorithms work directly on the dataset
  - They need to be extended to the whole space
- **So specifying a clustering algorithm should mean specifying**
  - How to label the sample points
  - How to extend this labeling to the whole space
- **Example: use 1-NN to build a partition of the space**
- **Extension is a prediction problem: the data and the partition are the training sample, the extended clustering is a model built by a multiclass learning algorithm.**
  - Yet another place for introducing bias

Outline

- Can we build a theory of clustering?
- What is clustering?
  - What is clustering?
  - What is the quality of a clustering?
  - What is the quality of an algorithm for a problem?
  - What is the quality of an algorithm in general?
- What is the right model for analyzing it?
- What are the right questions?
- Conclusion

pertinence
simple solutions for complex processes

Quality assessment

- Quality measures of a clutering: measure the “fit”
- Quality measures of an algorithm
  - not geared towards a sample but towards the distribution
  - can be estimated on a sample
  - aimed at selecting the “right” number of clusters
- Examples
  - Penalty term (BIC, MDL): arbitrary choice
  - Cross-validated quality measure: requires extension (hence prediction)
  - Gap statistic
  - Stability

pertinence
simple solutions for complex processes

Stability

- **"If the clustering is stable (with respect to small changes in the dataset) it captures relevant structure"**
- **How to define it?**
  - Need to resample
  - Need to compare clusterings (essentially a prediction problem)
    - On the same set
    - On different sets (extension)
- **It cannot be defined without assumptions**

pertinence
simple solutions for complex processes

Stability and prediction

- **Lange et al. 2002**
  - Cluster the first half
  - Cluster the second half and extend it to the first half
  - Compare the labels
- **Ben-David & Schaefer 2005**
  - Cluster the first half
  - Cluster the second half
  - Compare both extensions (to the union)
- **Ben-David 2005**
  - Cluster S1 U S2
  - Cluster S1 U S3
  - Compare the labels on S1 (no need for extension, but S1 introduces a bias)
- **Quality measure based on prediction**
  - Cluster the whole data
  - Extend half of the labels to the other half
  - Compare the labels

pertinence
simple solutions for complex processes

Stability: supervised vs unsupervised

- **Distance between clusters or between losses?**
- **Supervised stability**
  - stable algorithms satisfy some sort of Lipschitz condition
  - need to assess variability in loss only (estimation error)
- **Unsupervised**
  - quality of a clustering is not only measured by a "quality measure" (no unique goal)
  - assumption: for large sample sizes, the clustering has converged to an "optimal one", stability measures how far we are from this

pertinence
simple solutions for complex processes

Issues with stability

- **Does not capture the "fit" (Iris example in Lange et al 2002)**
  - Need to avoid stable but trivial solutions
- **Stability measures several effects**
  - Sampling sensitivity of the algorithm
  - Degeneracy of the quality measure (different clusterings of the same sample may have the same quality)
  - Stability of the algorithm itself (for stochastic algorithms, even for a fixed sample, there may be different clusterings)
- **One needs to consider the distribution of possible clusterings when the sample is perturbed (Buhmann).**
- **Making a clustering of them allows to avoid the degeneracy issue, but introduces yet another bias.**

pertinence
simple solutions for complex processes

Outline

- Can we build a theory of clustering?
- What is clustering?
  - What is clustering?
  - What is the quality of a clustering?
  - What is the quality of an algorithm for a problem?
  - What is the quality of an algorithm in general?
- What is the right model for analyzing it?
- What are the right questions?
- Conclusion

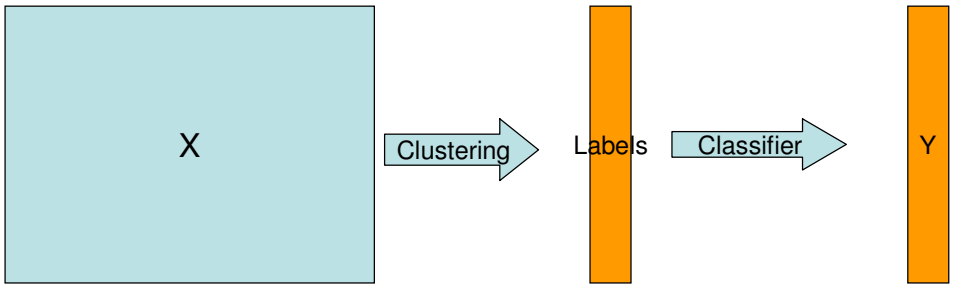
pertinence
simple solutions for complex processes

Evaluation on classification problems

- Goal: estimate the quality of an algorithm in general or for a class of problems
  - Consider a classification problem
  - Determine if clustering helps for classification
- Can be used as a benchmark but this should not be the goal of clustering

pertinence
simple solutions for complex processes

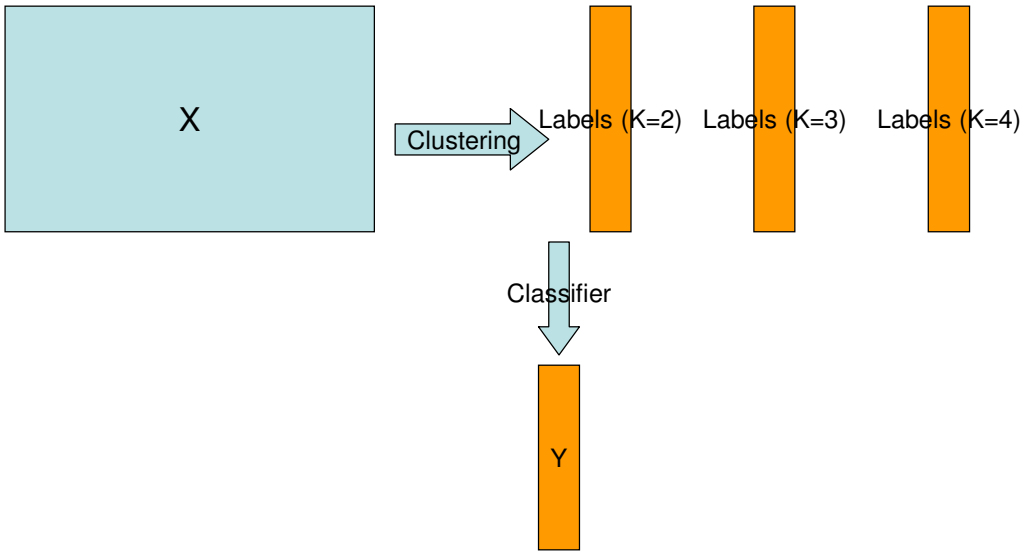
Option 1: Banerjee & Langford



- **Use a simple learning algorithm to compare labels**

pertinence

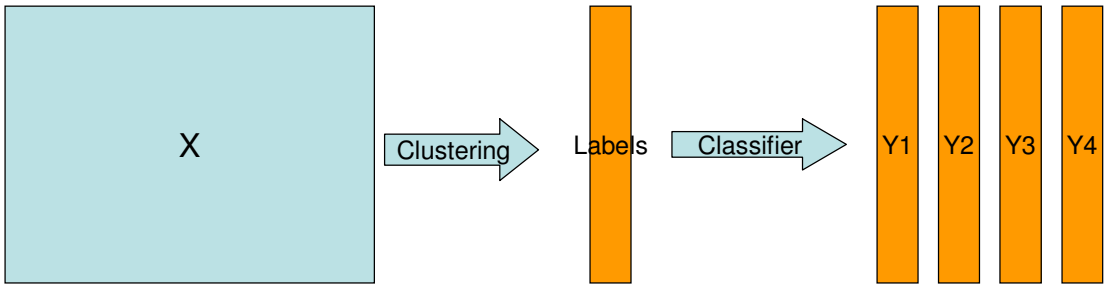
Option 2: Candillier et al.



- **Build new features from clusterings at different scales**
- **Use these new features for predicting Y**

pertinence

Option 3



- Consider several classification problems from the same input data (e.g. predict author and topic from texts)
- Average error over those

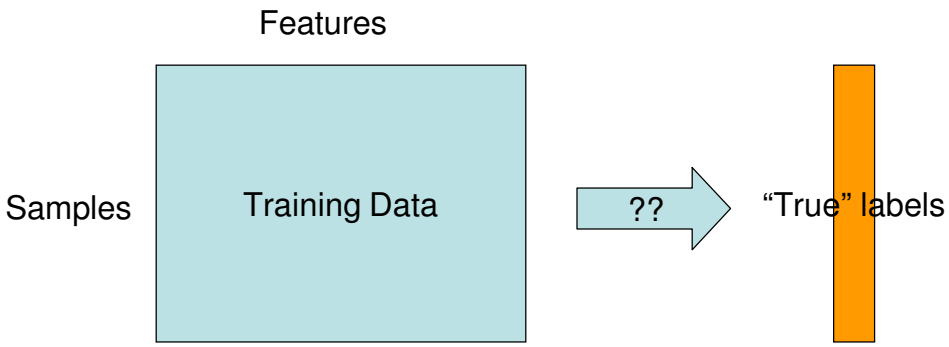
pertinence

Outline

- Can we build a theory of clustering?
- What is clustering?
- What is the right model for analyzing it?
- What are the right questions?
- Conclusion

pertinence

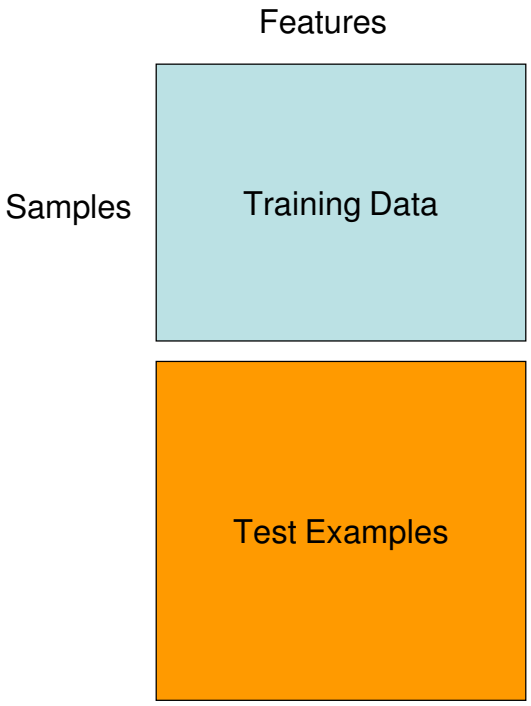
What is the right model ?



- **How do we model the relationship between samples and the labels that we have to find?**
- **We need to make assumptions on the way the data is generated**

pertinence

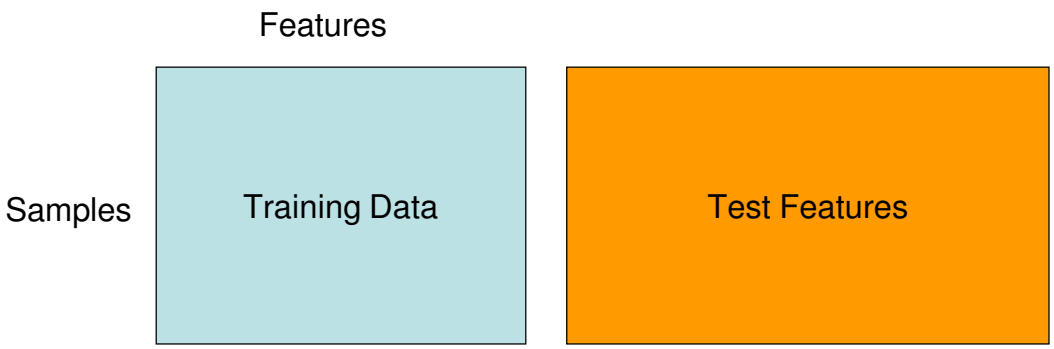
Model 1: sampling of examples



- **Assume examples have been randomly sampled (iid)**
- **Generalization means uncovering features of the whole distribution**

pertinence

Model 2: sampling of features



- **Assume features have been randomly sampled (iid)**
- **Generalization means building labels that correlate well with unseen features**

pertinence

Outline

- **Can we build a theory of clustering?**
- **What is clustering?**
- **What is the right model for analyzing it?**
- **What are the right questions?**
- **Conclusion**

pertinence

Fundamental questions

- **Basic intuition: the more data you get, the more accurate the results are**
- **2 Questions (von Luxburg and Ben-David)**
- **Q1: Define the goal when the whole distribution is known**
  - This is a conceptual question which can be answered with a definition
  - this definition should satisfy some continuity with respect to the sampling
- **Q2: How to attain this from a finite sample?**
  - This is an algorithmic question which can be formally answered

pertinence
simple solutions for complex processes

Sub-questions

- **Estimation**
  - Convergence: Does the algorithm converge?
  - Rates: How fast does the algorithm converge?
- **Approximation**
  - Consistency: How good is the limit clustering?
  - Rates: How good is the clustering on a finite sample?
- **Computation**
  - How fast is the algorithm?
- **Examples**
  - Convergence and estimation rates for k-means (Pollard, Ben-David)
  - Convergence of Spectral Clustering (von Luxburg and B.)
  - Convergence and estimation rates in the feature sampling model (Krupka and Tishby)

pertinence
simple solutions for complex processes

> Model selection

- **Supervised learning**
  - Primary goal: estimate expected error
  - expected error = empirical error + corrective term
  - corrective term: variance estimate (cross-validation, complexity, or instability)
- **This can be applied directly to clustering, if the goal is just to find the model with smallest expected quality**
- **However, quality is usually not sufficient to determine K unambiguously (degeneracy problem)**
  - Is the expected quality a good measure?
  - Does the quality measure have a unique optimum?
    - If not, are they all for the same K?
    - Is this optimum attained for a finite K?
  - Is there a value of K that can be agreed upon? (consider the case of groups of clusters)

> Outline

- **Can we build a theory of clustering?**
- **What is clustering?**
- **What is the right model for analyzing it?**
- **What are the right questions?**
- **Conclusion**

> Messages/Open Questions

- **Biases/Assumptions/Hypotheses cannot be avoided**
  - so make them explicit
  - Instead of hiding the bias, try to find principled ways of incorporating or choosing it.
  - A convenient way is to use a classification algorithm
- **Unsupervised theory cannot do more than supervised theory**
  - it cannot tell you which algorithm is better
  - it can only answer specific and partial questions once the definitions have been set
- **The question of what is the right model is still open**

- **Stability**
  - Bias-variance trade-off: how to measure and trade both?
  - Degeneracy of the objective and other sources of instability (can regularization help?)