1.4 Advanced Issues in Data Preparation and Modeling

Data Reduction

Context

Goals

- obtain a reduced representation of the data set that is much smaller in volume, producing the same analytical results (or almost the same)
- improved visualization of data with more interpretable models
- much faster

Dimensionality Reduction

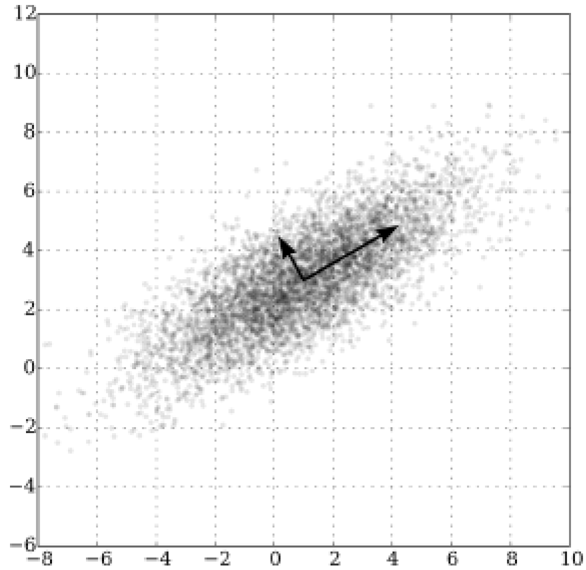
- Dimensionality increases:
    - data becomes increasingly sparse
    - density and distance between points becomes less meaningful
    - possible combinations of subspaces will grow exponentially
- number of data points required for robust patterns grows exponentially with number of attributes
- **2 Approaches**: Attribute Aggregation, Feature Selection

Attribute Aggregation

Principal Component Analysis (PCA)

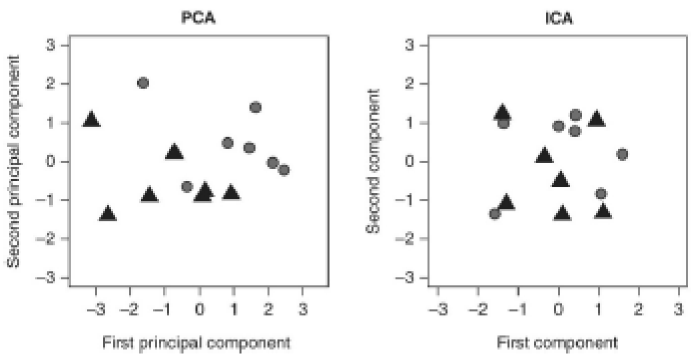
- n new features
- linear combinations of existing n attributes
- orthogonal to each other

- k << n explain most of the variance in the data



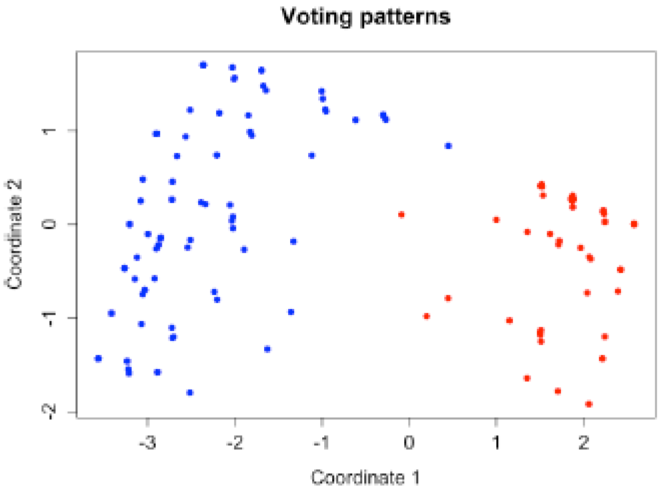
ICA vs. PCA

- Both create linear combinations of the attributes
- **ICA**
  - assumes the original attributes are statically independent
  - reduces higher order statistics
  - does not rank components



Multidimensional Scaling

- linear projection of a data set
- uses the distances between pairs of objects
- particularly suitable when it is difficult to extract relevant features to represent the objects



Feature Selection

- **Eliminate**:
  - **redundant attributes**: duplicate much or all of the information contained in one or more other attributes
  - **irrelevant attributes**: contain no useful information

Filter Methods

- 2 attributes: remove redundant attributes
- 1 attribute vs. target: identify relevant variables

Data Modeling

Context

	Yes	**No**
Yes	TP = True Positive	FN = False Negative
No	FP = False Positive	TN = True Negative

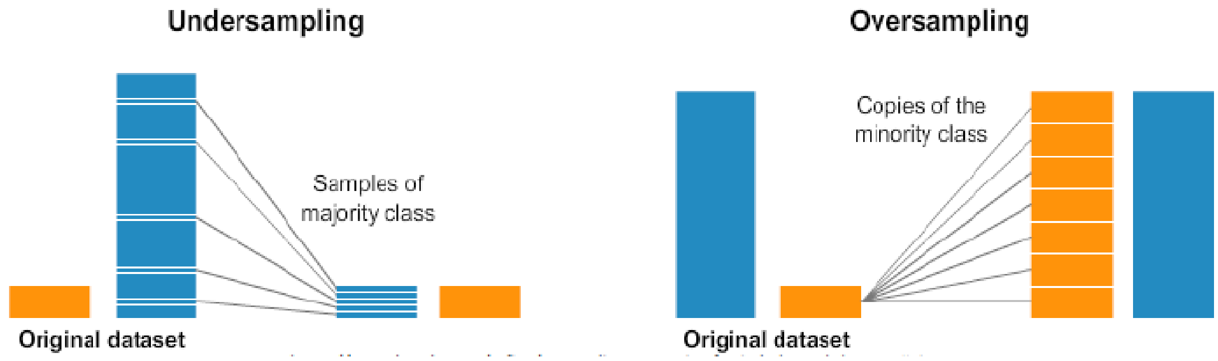
- ML methods usually minimize FP + FN, but potentially FP >> FN, so algorithm effectively minimizes FP

Class Imbalance

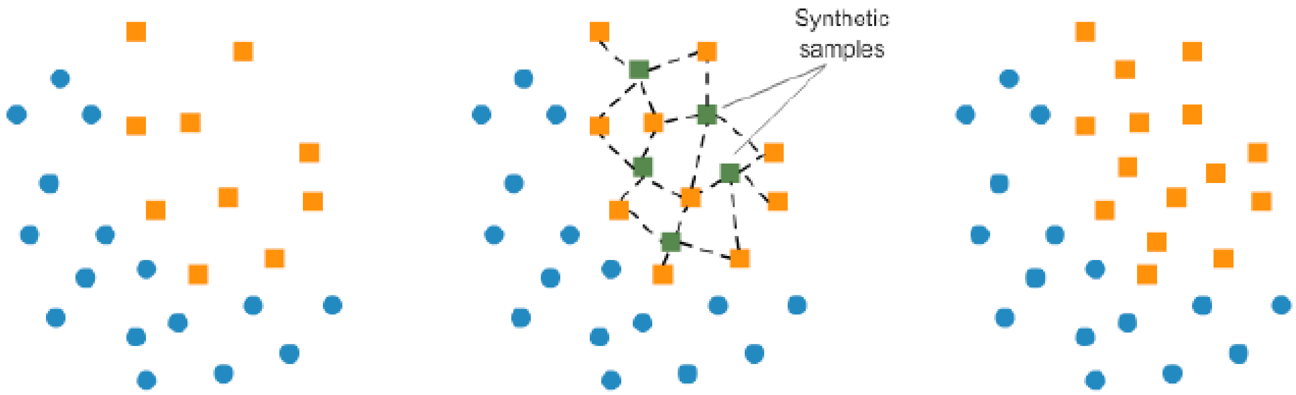
- collect more data
- resample existing data
- create synthetic data (e.g. SMOTE)
- adapt your learning algorithm (e.g. cost sensitive learning)

Resampling

- **Undersampling**: possible loss of information
- **Oversampling**: fixed boundaries and danger of overfitting



SMOTE (Synthethic Minority Over-sampling Technique)



- Possibility of inadequate boundaries and danger of overfitting

Cost Sensitive Learning

- FP and FN error often incur different costs, but ML methods still usually minimize FP+FN
- Simple methods:
    - resampling according to costs
    - weighting according to costs (basically, the same thing)
- Complex methods: e.g. metacost

Metacost

- independent of algorithm

1. create bootstrap replicates of training data
2. learn model from each replicate
3. relabel examples

$$argmin_i \sum_j P(j|x)C(i|j)$$

- $C(i|j)$ = cost of mistaking j by i
- $P(j|x)$ = class probability of x by voting

4. learn model on relabelled data

Data Quality: multidimensional view

- accuracy
- completeness
- consistency
- timeliness
- believability
- interpretability

Wrong reasons for believing that data is clean

- data warehouse
- IS was just revamped
- major data cleanup
- data collected automatically
- data collection is human-error proof
- "tell us what you need: we have everything"

How to do better?

- Human resources
- analytics at the core of IS development
- data quality is a continuous process

Data Cleaning as a Process

1. Discrepancy detection
    - validate with metadata
    - check field overloading
    - check uniqueness rule, consecutive rule and null rule
    - commercial tools (scrubbing and auditing)
2. Migration and Integration
    - data migration tools
    - ETL (Extraction/Transformation/Loading) tools

Automation

- automl & metalearning - some progress on algorithm selection, early work on workflow, not really data cleaning

DQaaS

- Yes, it automation is possible

- Issues: confidentiality, computational costs