SEKOLAH TINGGI MANAJEMEN INFORMATIKA DAN TEKNIK KOMPUTER

STMIK STIKOM Bali

Data Mining Evaluatio

Made Satria Wibawa

Campus of Technology and Creativity

Outline

1. Kriteria evaluasi pendahuluan
2. Evaluasi empiris pengklasifikasi
    - Tahan
    - Validasi silang
    - Membiarkan satu dan teknik lainnya
3. Skema lainnya

Classification Problem

• Tugas umum: menugaskan label kelas keputusan ke satu set objek yang tidak diklasifikasikan yang dijelaskan oleh seperangkat atribut (fitur) tetap.
• Dengan seperangkat contoh pra-klasifikasi, temukan klasifikasi representasi pengetahuan,
• untuk digunakan sebagai penggolong untuk mengklasifikasikan kasus baru (perspektif prediktif) atau untuk menggambarkan situasi klasifikasi dalam data (perspektif deskriptif).
• Pembelajaran yang diawasi: kelas dikenal dengan contoh yang digunakan untuk membangun classifier.

Approaches to learn classifiers

- Decision Trees
- Rule Approaches
- Logical statements (ILP)
- Bayesian Classifiers
- Neural Networks
- Discriminant Analysis
- Support Vector Machines
- k-nearest neighbor classifiers
- Logistic regression
- Artificial Neural Networks
- Genetic Classifier
- ……

Discovering and evaluating classification knowledge

Creating classifiers is a multi-step approach:

- Generating a classifier from the given learning data set,
- Evaluation on the test examples,
- Using for new examples.

Train and test paradigm!

Evaluation Criteria (1)

• Prediktif (Klasifikasi) akurasi: ini mengacu pada
Kemampuan model untuk memprediksi kelas dengan benar
label data baru atau yang sebelumnya tidak terlihat:
ketepatan =% contoh uji coba diklasifikasikan dengan benar
oleh pengklasifikasi

• Kecepatan: ini mengacu pada biaya komputasi yang
terlibat
dalam menghasilkan dan menggunakan model

• Kekokohan: inilah kemampuan model untuk dibuat
prediksi yang benar diberikan data bising atau data dengan
Nilai yang hilang

Evaluation Criteria (2)

• Skalabilitas: ini mengacu pada kemampuan untuk membangun
Model efisien diberikan sejumlah besar data
• Interpretabilitas: ini mengacu pada tingkat pemahaman dan wawasan yang diberikan oleh model
• Kesederhanaan:
ukuran pohon keputusan
aturan kekompakan
• Indikator kualitas yang bergantung pada domain

Evaluating Classifier

Predictive (classification) accuracy (0-1 loss function)

- Use testing examples, which do not belong to the learning set
  - N_t number of testing examples
  - N_c number of correctly classified testing examples
- Classification accuracy : $\eta = \frac{N_c}{N_t}$
- (Misclassification) Error : $\varepsilon = \frac{N_t - N_c}{N_t}$
- Other options:
  - analysis of confusion matrix

Binary Classification

	Predicted	
Original Classes	**YES**	**NO**
YES	**40**	**10**
NO	**5**	**45**

Confusion matrix :

True Positive (TP)

True Negative (TN)

False Positive (FP)

False Negative (FN)

Performance Parameter :

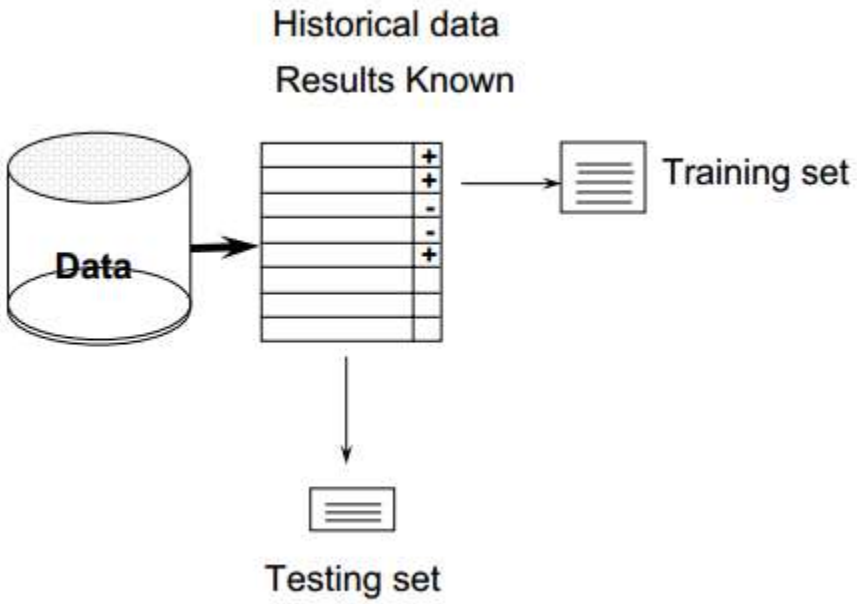
Accuracy : $\frac{(TP+TN)}{Total}$

Error Rate : $\frac{(FP+FN)}{Total}$

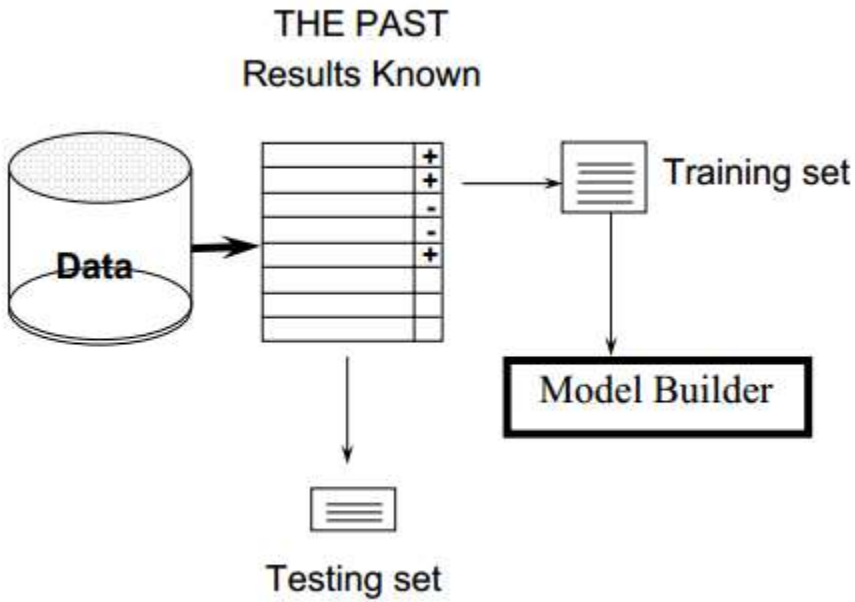
Precision : $\frac{TP}{(TP+FP)}$

Recall : $\frac{TP}{(TP+FN)}$

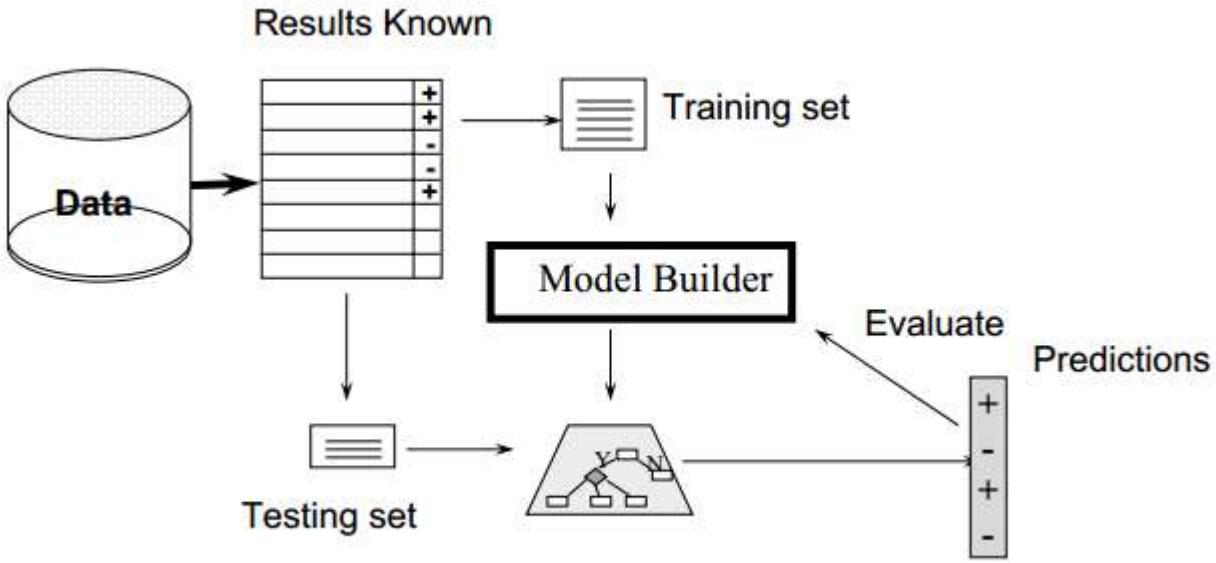
Step 1: Split data into train and test sets



Step 2: Build a model on a training set



Step 3: Evaluate on test set

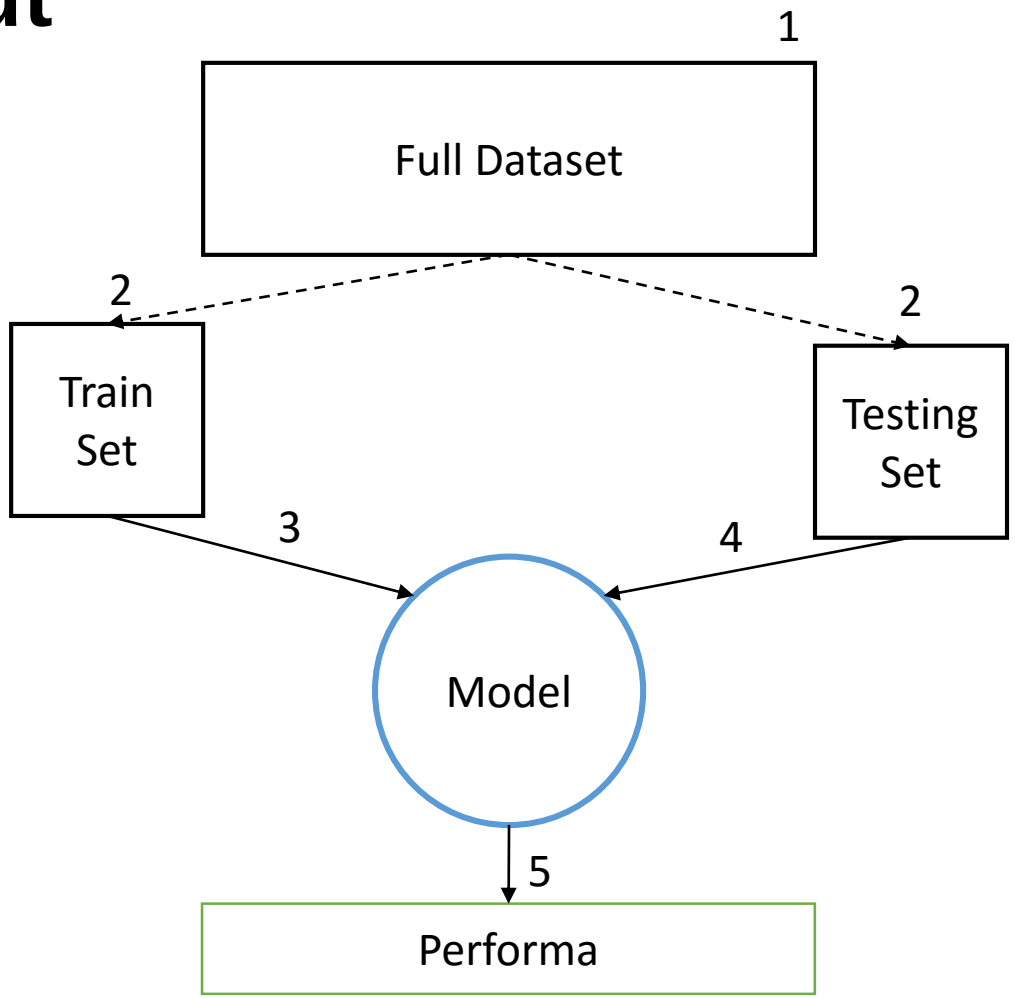


Experimental estimation of classification accuracy

Random partition into **train and test** parts:

1. Hold-out
    - use two independent data sets, e.g., training set (2/3), test set(1/3); random sampling
    - repeated hold-out
2. *k*-fold cross-validation
    - randomly divide the data set into *k* subsamples
    - use *k-1* subsamples as training data and one sub-sample as test data repeat *k* times
3. Leave-one-out for small size data

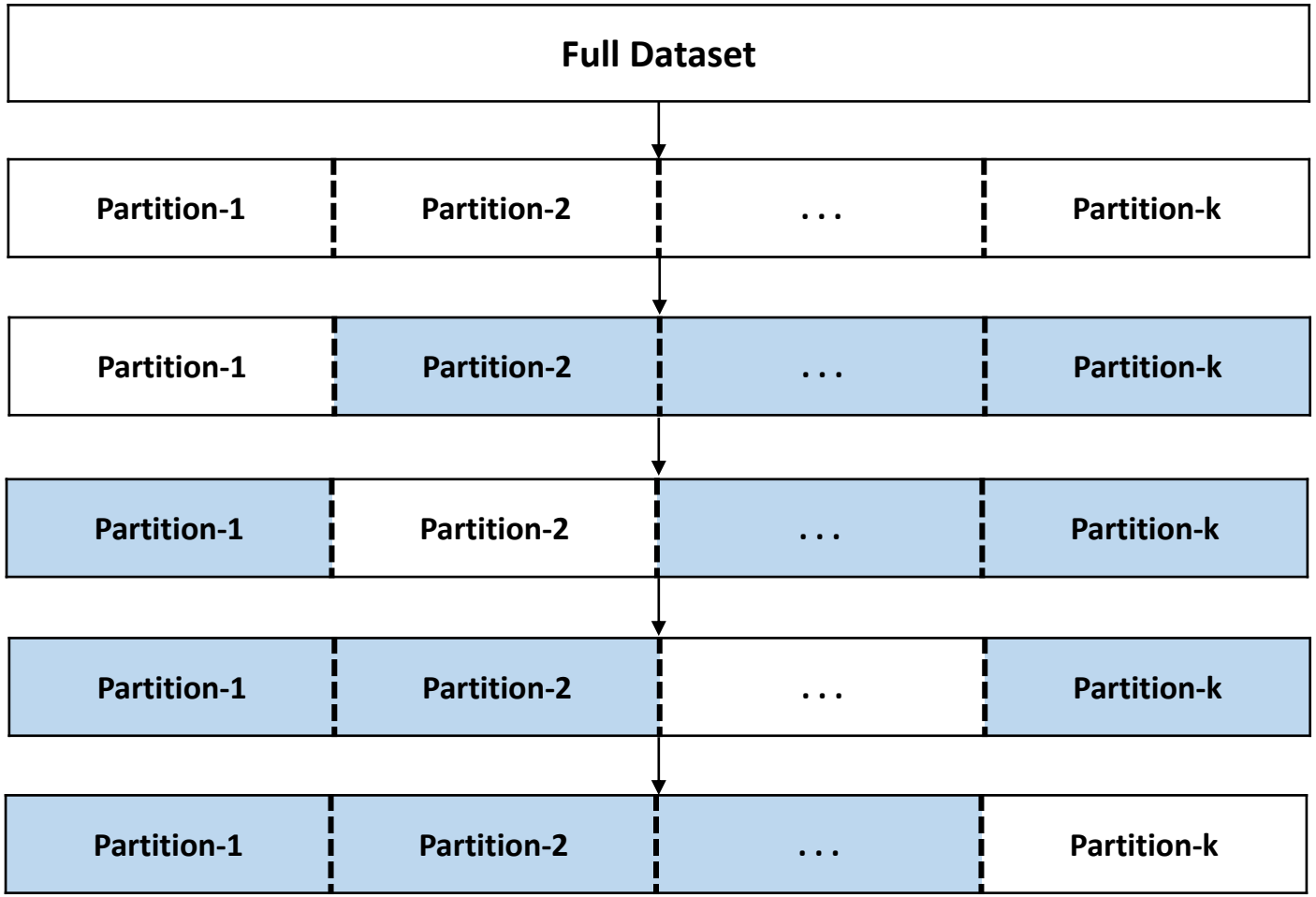
Hold out



Hold Out

- The simplest kind of cross validation.
- The data set is separated into two sets, called the training set and the testing set.
- The advantage of this method is that it is usually preferable to the residual method and takes no longer to compute.
- Its evaluation can have a high variance.
- The evaluation may depend heavily on which data points end up in the training set and which end up in the test set
- The evaluation may be significantly different depending on how the division is made.

k-Cross validation



k-Cross validation

One way to improve over the holdout method.

The data set is divided into *k* subsets, and testing and training is repeated *k* times. Each time, one of the *k* subsets is used as the test set and the other *k-1* subsets are put together to form a training set. Then the average error across all *k* trials is computed.

Advantages :

- The variance of the resulting estimate is reduced as *k* is increased.
- How the data gets divided is matters less
- Every data point gets to be in a test set exactly once, and gets to be in a training set *k-1* times.

Disadvantages :

- More computation time.

Leave-One-Out cross-validation

• Tinggalkan Satu Out: Suatu bentuk validasi silang tertentu. Jumlah lipatan dengan jumlah instancesi pelatihan., Untuk n contoh pelatihan, bangun classifier n kali tapi dari n - 1 contoh traning:

• Memanfaatkan sebaik-baiknya data

• Melibatkan tidak ada sub-sampling acak.