UNIVERSITAT DE BARCELONA

Course. Introduction to Machine Learning

Work 3. Lazy Learning exercise

Session 2

Dr. Maria Salamó Llorente

Dept. Mathematics and Informatics,

Faculty of Mathematics and Informatics,

University of Barcelona

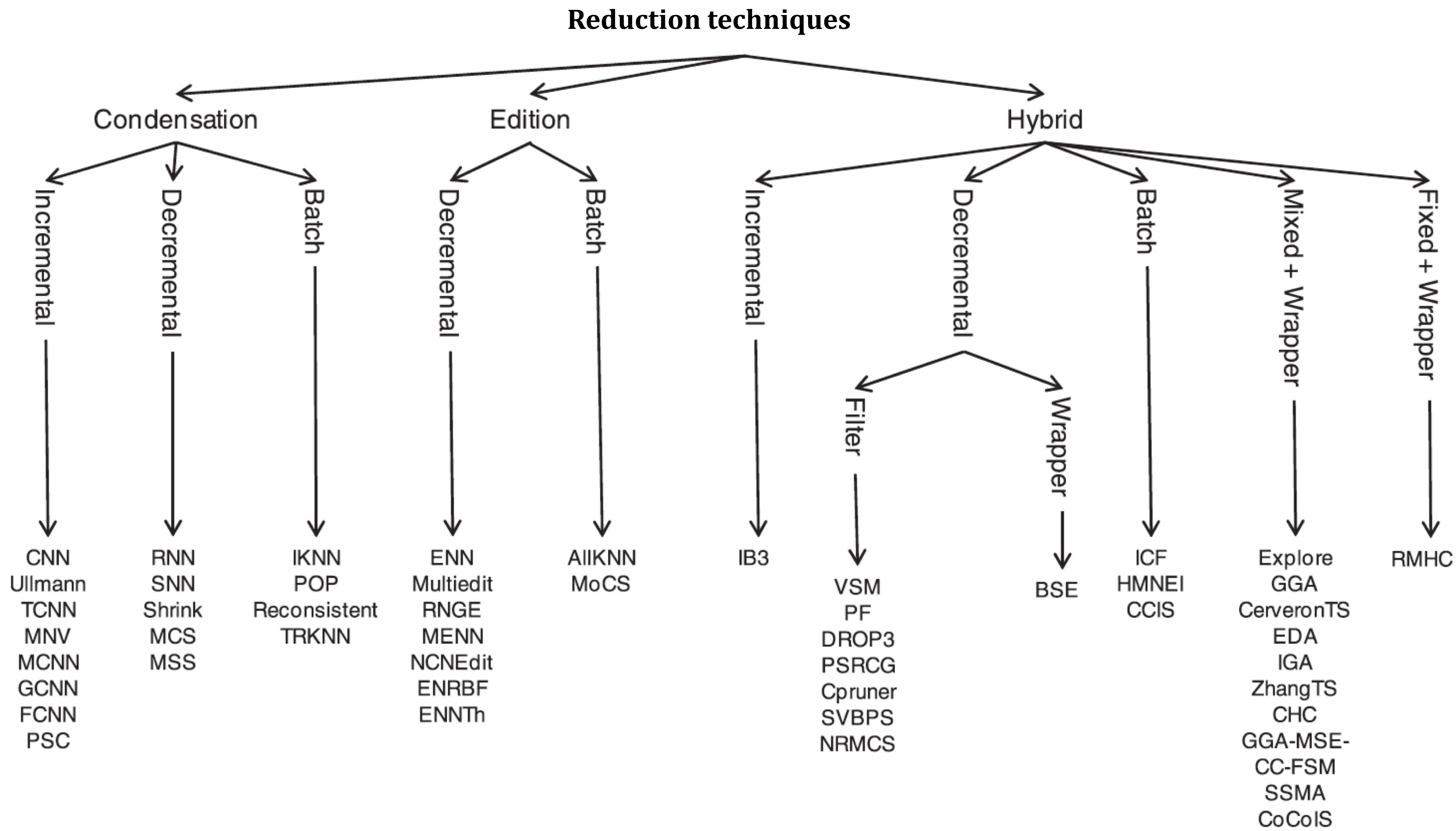
Contents

UNIVERSITAT DE BARCELONA

1. Introduction (Session 1)
2. Parameters in kNN (Session 1)
   1. Distance metric
   2. K parameter
   3. Voting schemes
   4. Weighting
3. Instance Reduction techniques (Session 2)

Instance reduction techniques

Taxonomy


Reduction techniques
Condensation
Edition
Hybrid
Incremental
Decremental
Batch
Decremental
Batch
Incremental
Decremental
Batch
Mixed + Wrapper
Fixed + Wrapper
Filter
Wrapper
CNN
Ullmann
TCNN
MNV
MCNN
GCNN
FCNN
PSC
RNN
SNN
Shrink
MCS
MSS
IKNN
POP
Reconsistent
TRKNN
ENN
Multiedit
RNGE
MENN
NCNEdit
ENRBF
ENNTh
AllKNN
MoCS
IB3
VSM
PF
DROP3
PSRCG
Cpruner
SVBPS
NRMCS
BSE
ICF
HMNEI
CCIS
Explore
GGA
CerveronTS
EDA
IGA
ZhangTS
CHC
GGA-MSE-
CC-FSM
SSMA
CoCoIS
RMHC

Types of selection

- **Condensation**
  - Aim to retain points which are closer to the decision boundaries, also called border points
  - The idea behind these methods is to preserve the accuracy over the training set, but the generalization accuracy over the test set can be negatively affected
  - Reduction rate is normally high
- **Edition**
  - Seek to remove border points
  - Remove points that are noisy or do not agree with their neighbors
  - Reduction rate is low
- **Hybrid**
  - Try to find the smallest subset S which maintains or even increases the generalization accuracy in test data
  - It allows the removal of internal and border points

Direction of search

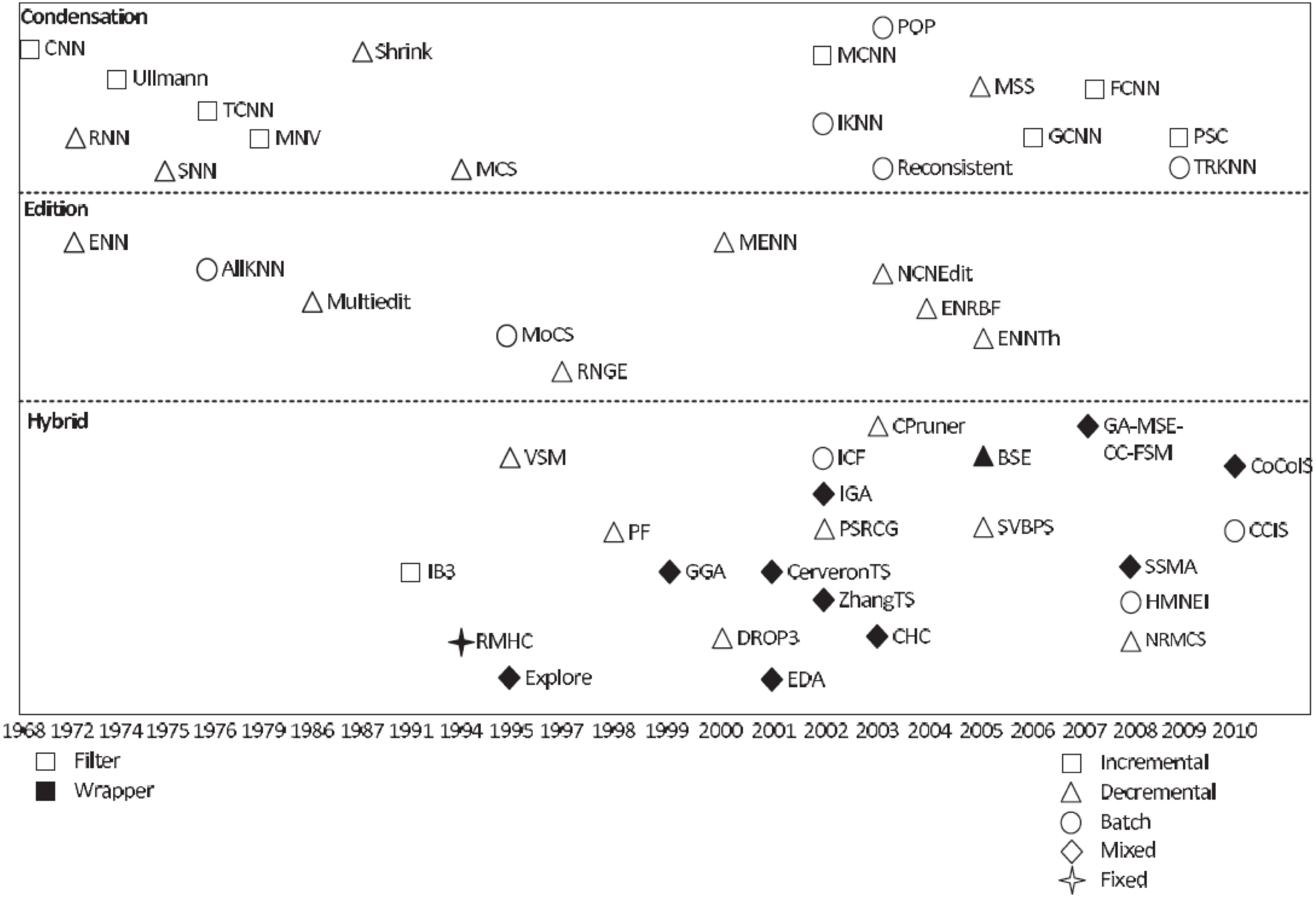
- **Incremental**
  - Starts with an empty subset S, and adds each case in the training set to S if it fulfills some criteria
  - These algorithms depend on the order of presentation
- **Decremental**
  - Begins with S= training set, and then searches for instances to remove from S
  - These algorithms depend on the order of presentation
- **Batch**
  - Use a batch mode. This involves deciding if each instance meets the removal criteria before removing any of them.
  - Then, all those that do not meet the criteria are removed at once
- **Mixed**
  - Begins with a preselected subset S (randomly or selected by an incremental or decremental process) and can iteratively add or remove any instance which meets the specific criterion

Evaluation of search

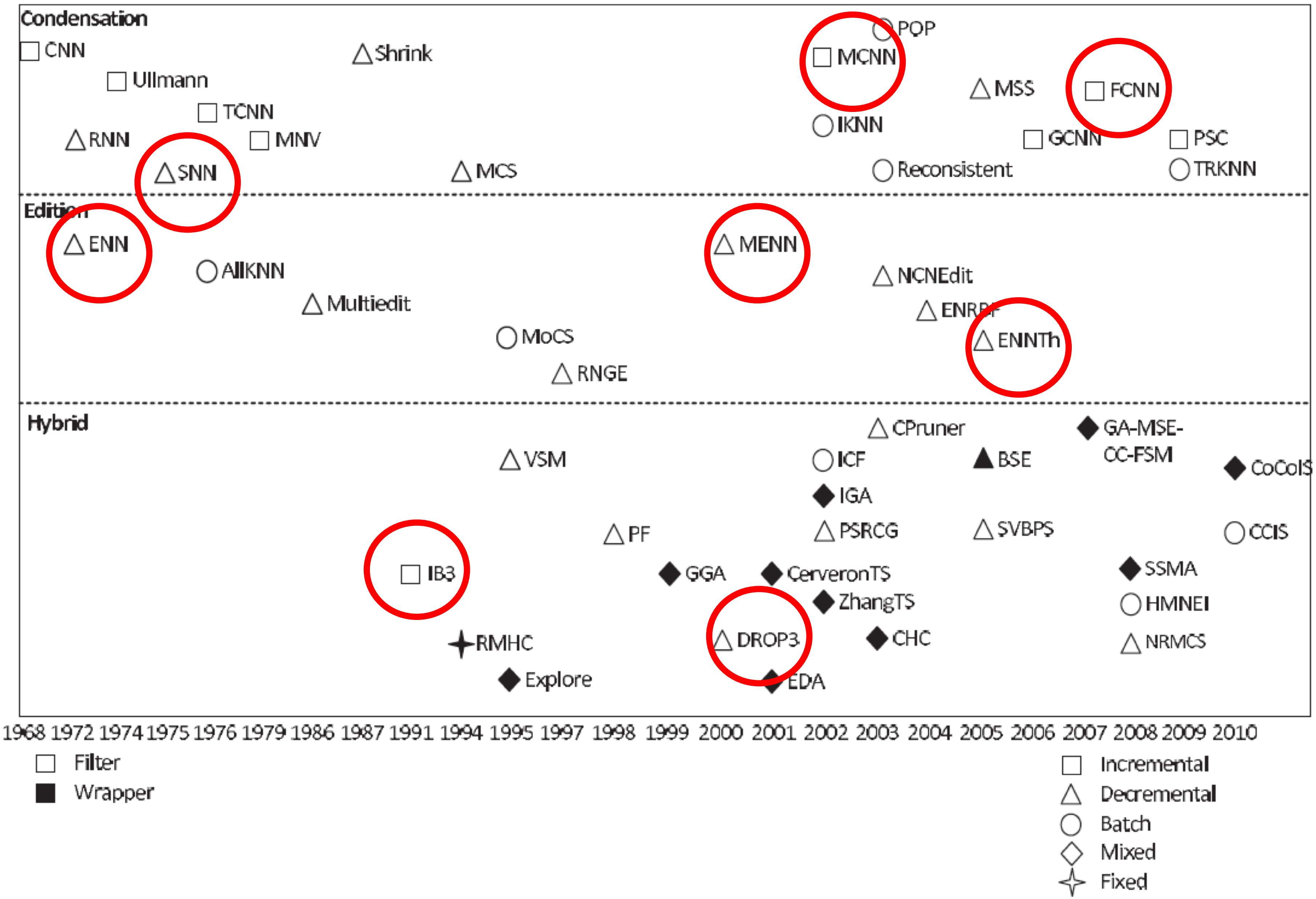
- **Filter**
  - When the kNN rule is used for partial data to determine the criteria of adding or removing and no leave-one-out validation scheme is used to obtain a good estimation of generalization accuracy
- **Wrapper**
  - When the kNN rule is used for the complete training set with the leave-one-out validation scheme
  - The conjunction in the use of the two mentioned factors allows us to get a great estimator of generalization accuracy, which helps to obtain better accuracy over test data
  - It can be computationally expensive

UNIVERSITAT DE BARCELONA

Reduction techniques map



Reduction techniques map



Criteria to compare methods

- **Storage reduction**
  - The main goal of reduction techniques is to reduce storage requirements
  - Another goal is to speed up classification
- **Noise tolerance**
  - Two main problems may occur in the presence of noise:
    1. Few instances will be removed because many instances are needed to maintain noisy decision boundaries
    2. The generalization accuracy can suffer, especially if noisy instances are retained instead of good instances
- **Generalization accuracy**
  - A successful algorithm will be often able to significantly reduce the size of the training set without significantly reducing the accuracy
- **Time requirements**
  - If the learning phase takes too long it can become impractical for real applications

Reduction techniques in Work 3

- **SNN or FCNN or MCNN**
  - **Filter** approaches based on **condensation** with **incremental** or **decremental** direction of search

  - SNN: G.L. Ritter, H.B. Woodruff, S.R. Lowry, and T.L. Isenhour, “An Algorithm for a Selective Nearest Neighbor Decision Rule,” IEEE Trans. Information Theory, vol. 21, no. 6, pp. 665-669, Nov. 1975
  - FCCN: F. Angiulli, “Fast Nearest Neighbor Condensation for Large Data Sets Classification,” IEEE Trans. Knowledge and Data Eng., vol. 19, no. 11, pp. 1450-1464, Nov. 2007
  - MCNN: V.S. Devi and M.N. Murty, “An Incremental Prototype Set Building Technique,” Pattern Recognition, vol. 35, no. 2, pp. 505-513, 2002

CNN

- **CNN Family**
  - ***Condensed Nearest-Neighbor rule* (CNN)**
    - build an edited set from scratch by adding instances that cannot be successfully solved by the edited set built so far
    - tends to select training instances near the class boundaries.
    - consistent
    - not minimal edited set (redundant instances) : order dependent
  - ***Reduced Nearest-Neighbor* (RNN) method**
    - adaptation of CNN
    - postprocess to contract the edited set by identifying and deleting redundant instances

UNIVERSITAT DE BARCELONA

Reduction techniques in Work 3

- **ENN or Modified ENN or ENNTh**
  - **Filter** approaches based on **edition** with **decremental** direction of search
  - **ENN**: D.L. Wilson, “Asymptotic Properties of Nearest Neighbor Rules Using Edited Data,” IEEE Trans. Systems, Man, and Cybernetics, vol. 2, no. 3, pp. 408-421, July 1972
  - **MENN**: K. Hattori and M. Takahashi, “A New Edited K-Nearest Neighbor Rule in the Pattern Classification Problem,” Pattern Recognition, vol. 33, no. 3, pp. 521-528, 2000
  - **ENNTh**: F. Vázquez, J.S. Sánchez, and F. Pla, “A Stochastic Approach to Wilson’s Editing Algorithm,” Proc. Second Iberian Conf. Pattern Recognition and Image Analysis, pp. 35-42, 2005

ENN

- **Edited Nearest Neighbor**
  - perfect counterpoint to CNN
  - filter out incorrectly classified instances in order to remove boundary instances (and noise) and preserve interior instances that are representative of the class being considered

Procedure

  - begin with all training instances
  - removed if its classification is not the same as the majority classification of its *k* nearest neighbors (edits out the noisy and boundary instances)
  - suffer from redundancy problem

Reduction techniques in Work 3

- IB2 or IB3 or DROP2 or DROP3
  - **Filter** hybrid approaches
  - **IB2 or IB3**: D.W. Aha, D. Kibler, and M.K. Albert, "Instance-Based Learning Algorithms," Machine Learning, vol. 6, no. 1, pp. 37-66, 1991
  - **DROP2 or DROP3**: D.R. Wilson and T.R. Martinez, "Reduction Techniques for Instance-Based Learning Algorithms," Machine Learning, vol. 38, no. 3, pp. 257-286, 2000

IBL family

- **IBL (Instance Based Learning) Family**
  - **IB1**
    - similar to CNN
  - **IB2**
    - makes one pass -> does not guarantee consistency
    - suffer from redundancy and sensitive to noisy data
  - **IB3**
    - reduce the noise sensitivity by only retaining *acceptable* misclassified instances
    - record for each instance which keep track of the number of correct and incorrect classifications
    - significance test : good classifiers are kept

IB-Algorithms

- **IB1**: store all examples
    - High noise tolerance
    - High memory demands
- **IB2**: Store examples that are misclassified by current example set
    - Low noise tolerance
    - Low memory demands
- **IB3**: like IB2 but,
    - Maintain a counter for the number of times the example participated in correct and incorrect classifications
    - Use a significant test for filtering noisy examples
        - Improved noise tolerance
        - Low memory demands

IB2

- IB2 is an extension to IB1 algorithm
  - Save memory and speed up classification
  - Unnecessary to use all data points for classification
- Algorithm
  - Work with data points incrementally
  - For each newly received data point apply NN using already saved points to predict its class
  - Only remember misclassified instances for future predictions
  - Problem:
    - Important instances in the early moments of learning are discarded
    - Noisy data gets incorporated

IB3

- IB3 is an extension of IB1
  - Deal with noise, keep only good classifier data points
  - Discard instances that do not perform well
- Algorithm:
  - Keep a record of the number of correct and incorrect classification decisions that each saved data point makes
  - Two predetermined thresholds are set on success ratio
  - An instance is selected to be used for training:
    - If the number of incorrect classifications is ≤ the first (lower) threshold and,
    - If the number of correct classifications is ≥ the second (upper) threshold

Drop Family

- **Drop Family**
  - guided by two sets for each instances : *k NN*s & *associates* of instance
  - associates of *i* : those cases which have *i* as one of their nearest neighbors
  - begin with the entire training set
  - *i* is removed if at least as many of its associates can be correctly classified without *i*
- **Drop1**: tends to remove noise from the original case-base
- **Drop2**: cases are sorted in descending order of NUM distance
- **Drop3:** combines ENN pre-processing with DROP2 to remove noise and it is one of the best instance based classifier

UNIVERSITAT DE BARCELONA

References

- Janez Demšar. 2006. Statistical Comparisons of Classifiers over Multiple Data Sets. J. Mach. Learn. Res. 7 (December 2006), 1-30 (**MANDATORY READING**)

- Wilson, D.R., Martínez, T.R., 1997. Improved heterogeneous distance functions. Journal of Artificial Intelligence Research 6, 1–34
  - **This paper details the basis of the CNN family, ENN family, IB family and drop family**

- David W. Aha, Dennis Kibler, and Marc K. Albert. 1991. Instance-Based Learning Algorithms. Machine Learning. 6, 1 (January 1991), 37-66

- R. Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. In Proceedings of the International Joint Conferences on Artificial Intelligence IJCAI-95. 1995

UNIVERSITAT DE BARCELONA

Course. Introduction to Machine Learning

Work 3. Lazy Learning exercise

Session 2

Dr. Maria Salamó Llorente

Dept. Mathematics and Informatics,

Faculty of Mathematics and Informatics,

University of Barcelona