UNIVERSITAT DE BARCELONA

Course. Introduction to Machine Learning
Introduction to Case-Base Maintenance

Maria Salamó Llorente

Dept. Matemàtica Aplicada i Anàlisi (MAiA), Facultat de Matemàtiques
Universitat de Barcelona (UB)

Contents

Introduction to CBR

- *Case-based reasoning (CBR)*

Solves problems by reusing the solutions to similar problems stored as cases in a case-base

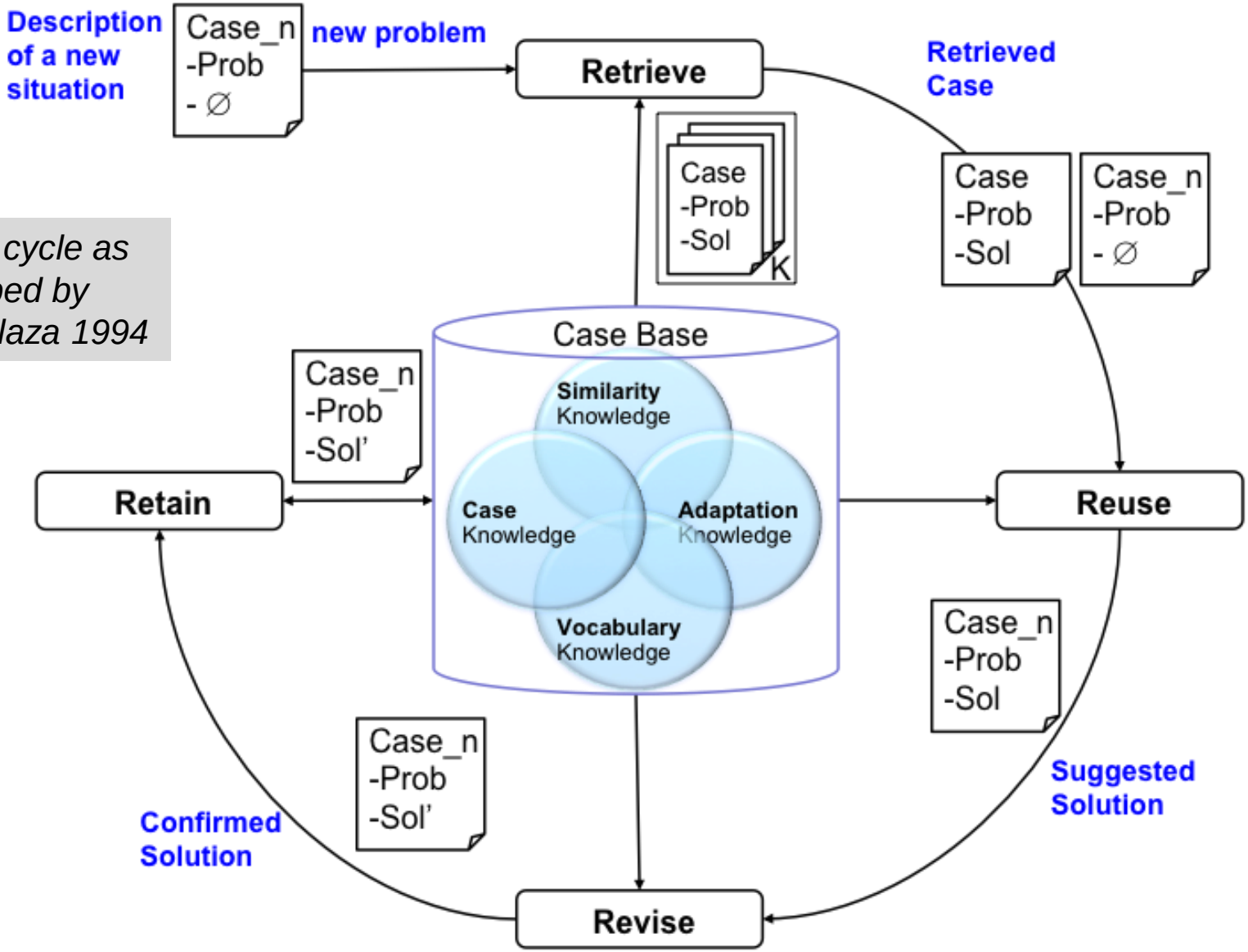
- *Footprint of a case-base:*

A minimal set of cases which is representative of the entire case-base

UNIVERSITAT DE BARCELONA

Introduction to CBR cycle

The CBR cycle as described by Aamodt&Plaza 1994



Case-Base Maintenance

- **Edited Instance Set**
  - NN classification algorithm suffers,
    - Large storage & computational costs
  - approach for reducing costs
    - Instance selection (editing technique)
  - properties of edited set
    1. **Size** : as few instances as possible
    2. **Consistency** : capable of correctly classifying all of the instances in the training set
    3. **Competency** : capable of correctly classifying unseen instances

CBM: 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

CBM: CNN

- CNN-NUN
  - NUN (nearest unlike neighbor)

    : distance to an instance's nearest neighbor in an opposing class
  - preprocess : ascending NUN distance
  - still suffer
  - s from noise problems
- problems of CNN family
  - do not always generalize well to unseen target instances
  - sensitive to noisy data

CBM: 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

CBM: RENN and All-kNN

- **RENN (repeated ENN)**
  - repeatedly applying ENN until all instances have the majority classification of their neighbors
  - the effect of widening the gap betwn classes and smoothing the decision boundaries
- **All-kNN**
  - increases the value of k for each iteration of RENN
  - the effect of removing boundary instances and preserving interior

CBM: IBL

- **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

CBM: Drop Family

- **Drop Family**
  - guided by two sets for each instances : *k NNs & 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

CBM: A Competence Model for CBR

- Foundations of Competence
  - *coverage set* of a *case*
    - the set of all *target problems* that this case can be used to solve
  - *reachability set* of a target probem
    - the set of all *cases* that can be used to solve it

$$CoverageSet(c \in C) = \{c' \in C : Solves(c, c')\}$$

$$\text{Re}achability\,Set(c \in C) = \{c' \in C : Solves(c', c)\}$$

CBM: A Competence Model for CBR

- Competence Groups
  - coverage and reachability sets provide a measure of local competence only

$$\text{Re}\,latedSet\,(c) = CoverageSet(c) \cup \text{Re}\,chabilityS\,et\,(c)$$

$$\text{For}\; c1, c2 \in C, SharedCoverage(c1, c2) \;\; iff$$

$$[\text{Re}\,latedSet\,(c1) \cap RElatedSet\,(c2)] \neq \{\}$$

$$\text{For}\; G = \{c1, .cn\} \subseteq C, CompetenceGroup(G) \;\; iff$$

$$\forall ci \in G, \exists cj \in G - \{ci\} : SharedCoverage(ci, cj)$$

$$\wedge \forall ck \in C - G, \neg \exists cl \in G : SharedCoverage(ck, cl)$$

  - competence group : maximal sets of cases exhibiting shared coverage
  - each case within a given competence group must share coverage with at least one other case in that group
  - no case from one group can share coverage with any case from another group

Competence Footprinting

- **competence footprint of a case-base**: small, consistent and highly competent subset
- **competence footprint of a case-base** : the union of the competence footprints over all competence groups
- **CNN Footprinting**
  - apply the basic CNN algorithm to each competence group in turn
  - footprint set of cases is produced by combining the cases selected from each competence group
  - depends on the presentation order of the competence group cases
  - tends to preserver some redundant cases

Competence Footprinting

- **RC Footprinting**
  - *relative coverage* (RC): estimates the unique competence contribution of an individual case

$$\mathrm{Re}\,lativeCoverage(c) = \sum_{c' \in CoverageSet(c)} \frac{1}{|\,\mathrm{Re}\,achability\,Set(c')}$$

if a case c' is covered by n other cases then each of the n cases will receive a contribution of 1/n from c' to their relative coverage measures

  - cases are arranged in descending order of RC prior to the CNN-FP footprinting procedure

Competence Footprinting

- **RFC Footprinting**
  - Idea : cases with small reachability sets are interesting because they represent problems that are difficult to solve
  - ascending order of their reachability set size
  - tends to select those cases that lie on the boundary of the competence group
- **Coverage Footprinting**
  - related to RFC-FP except that instead of biasing cases by their reachability set size, it biases cases by their coverage set size
  - decreasing order of their coverage set size
  - tends to adding internal cases from the competence groups