Course. Introduction to Machine Learning

Introduction to Instance Reduction

Maria Salamó Llorente

Dept. Mathematics and Informatics,

Faculty of Mathematics and Informatics,
University of Barcelona

Contents

Introduction to Instance Reduction

1. Introduction
2. Edited Instance Set
3. CNN family
4. Edited Nearest Neighbour
5. Instance-Based Learning family (IBL)
6. Drop family

Introduction

- Storing and using specific instances improves the performance of several supervised ML algorithms.
- Instance-based learning algorithms usually store all the training set but this causes:
  - A large storage is needed
  - The generalization process is slow
  - The data may contain inconsistencies and noise
- To deal with this, reduction techniques are used

Instance Reduction

- **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

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

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

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

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

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

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

Course. Introduction to Machine Learning

Introduction to Instance Reduction

Maria Salamó Llorente

Dept. Mathematics and Informatics,

Faculty of Mathematics and Informatics,
University of Barcelona