Abstract

- Building of classification models requires labeled examples
- Real-world data are often not labeled
- Human annotation/labeling process may be:
  - Time consuming & Costly
- Challenge:
  - Find ways of reducing the annotation/labeling effort
- Solution:
  - Learning with Soft-labels
    - Soft labels reflect the degree of belief of an annotator in the class labels
  - This work:
    - New method based on soft-label binning
    - Reduces the number of constraints

Introduction

- Soft labels:
  - Represent the degree of belief of an annotator in the class label using a probabilistic score
  - Can facilitate learning of classification models
- Caveat:
  - Humans are not good in providing consistent probabilistic assessments
  - Soft labels may be noisy
  - Learning methods should be robust to noisy soft labels

Related Work

- Existing works in literature show that there are two ways to learn with soft-label information:
  1. Regression on the exact soft labels \( p_i \)
  2. Respecting pairwise orderings rather than exact soft labels.
- Regressions
  - Idea:
    - Find a regression function \( f(x) \) that fits a logistic or linear regression model.
  - Disadvantage:
    - Does not work well when soft-labels are noisy
    - Respecting pairwise orderings rather than exact soft labels
  - Idea:
    - Find a discriminative projection \( f(x) \) that respects pairwise orderings of samples as much as possible
  - Disadvantage:
    - Number of constraints is \( O(N^2) \)

Reducing constraints via binning

- Reformulate as ordinal regression
  - Idea:
    - Solve jointly \( m \)-1 binary classifiers

Figures

- Figure 1. UCI Wine Red dataset (No noise)
- Figure 2. UCI Wine Red dataset (Weak noise)
- Figure 3. UCI Wine Red dataset (Strong noise)

Choosing the number of bins

- How to choose the number of bins?
  - \( N \) bins: equivalent to all pairwise constraints
  - 2 bins: equivalent to binary classification
  - The optimal number is somewhere between

- Heuristic function:
  - Inspired by optimal binning for discretization of continuous values
  \[#\) of \( bins = \text{floor}(\sqrt{N})\]

Experiments and Results

- We test our approach on both synthetic and real-world data.
- Noise generation:
  - Weak noise
  - Strong noise
- BinarySVM
  - Labels: Binary only
  - Model: Support vector machine
- SoftLogReg
  - Labels: Soft only
  - Model: Logistic regression
- GPR
  - Labels: Binary + Soft
  - Model: Gaussian process regression
- SoftSvmRankKNN
  - Labels: Binary + Soft
  - Model: SVM + pairwise constraints
- SoftSvmRankPair
  - Labels: Binary + Soft
  - Model: SVM + random pairwise constraints
- SoftSvmOrd-our method
  - Labels: Binary + Soft
  - Model: SVM + bin/sample constraints

Conclusion

- We develop a new robust method that uses soft-label information to reduce annotation effort.
  - Our method (1) can benefit greatly from soft-label information, and (2) is robust to different levels of soft-label noise.

Acknowledgements

The work presented in this paper was supported in part by grants R01GM088224 and R01LM010019 from the NIH. The content of the paper is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.