Improving Generalization by Data Categorization

Ling Li, Amrit Pratap, Hsuan-Tien Lin, and Yaser Abu-Mostafa

Learning Systems Group, Caltech

ECML/PKDD, October 4, 2005

Examples in Learning

A Learning System

Unknown Target $f \longrightarrow \boxed{\text{Examples } \{(\mathbf{x}_i, y_i)\}_i} \longrightarrow$ Learner

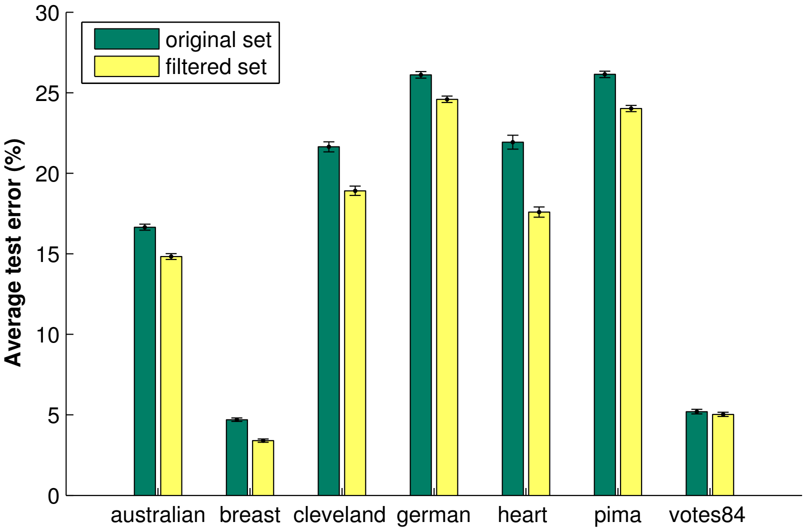
Examples are essential since they act as the information gateway between the target and the learner.

Not All Examples Are Equally Useful

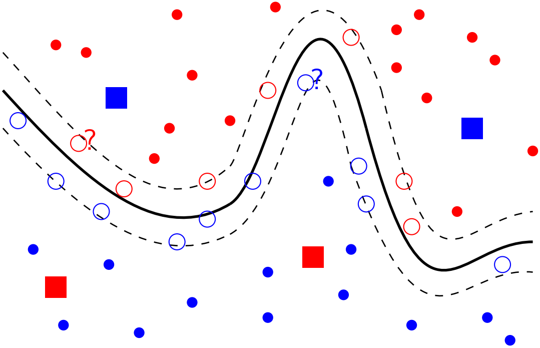
1. Surprising examples carry more information √
    - Garbage examples are also surprising (Guyon et al., 1996) ×
2. Noisy examples and outliers ×
3. Examples beyond the ability of the learner ×

Can we improve learning by automatically categorizing examples?

Improved Generalization



Categorize Examples

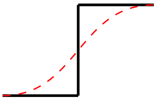


- Which examples are "bad"?
- Close-to-boundary examples are informative
- Three categories: typical, critical, and noisy

- The automatic data categorization is for better learning.
- The criteria are usually related with how useful or reliable the example is to learning, such as the margin.

Intrinsic Function

The target $f : \mathcal{X} \rightarrow \{-1, 1\}$ comes from thresholding an intrinsic function $f_r : \mathcal{X} \rightarrow \mathbb{R}$. That is



$$f(\mathbf{x}) = \operatorname{sign}(f_r(\mathbf{x})) .$$

Examples of $f_r(\mathbf{x})$

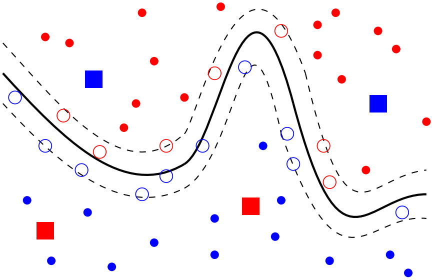
1. The credit score of the applicant $\mathbf{x}$ minus some threshold
2. The signed Euclidean distance of $\mathbf{x}$ to the boundary
3. The probability of $\mathbf{x}$ belonging to class 1 minus 0.5

Properties

- Problem-dependent (e.g., the knowledge of experts)
- Tells the usefulness or reliability of an example
- Unknown

Intrinsic Margin and Data Categorization

For an example $(\mathbf{x}, y)$, its intrinsic margin is $yf_r(\mathbf{x})$.

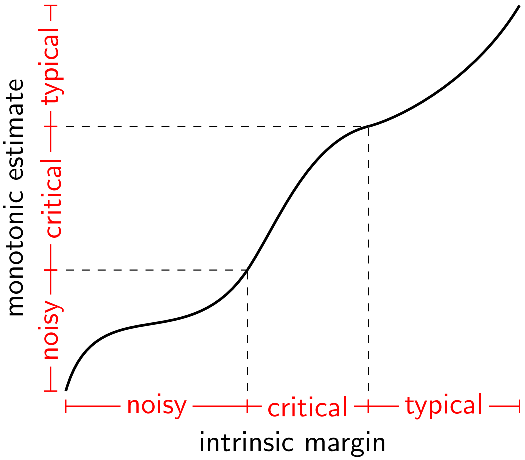


The intrinsic margin $yf_r(\mathbf{x})$ can be treated as a measure of how close $\mathbf{x}$ is to the decision boundary.

- Small positive: near the boundary — critical
- Large positive: deep in the class territory — typical
- Negative: mislabeled — noisy

Monotonic Estimate

However, the intrinsic margin is unknown.



- a monotonic estimate of the intrinsic margin
- two proper thresholds
- three categories

Selection Cost

For an example $(\mathbf{x}, y)$, a hypothesis g may classify it either wrongly or correctly. Consider the expected out-of-sample errors.

$$\mathrm{E}_g[\pi(g) \mid \underbrace{g(\mathbf{x}) \neq y}_{\text{wrongly}}] - \mathrm{E}_g[\pi(g) \mid \underbrace{g(\mathbf{x}) = y}_{\text{correctly}}]$$

We may select to trust $(\mathbf{x}, y)$, or not. The difference is the cost we pay when we make the selection. We call it the selection cost.

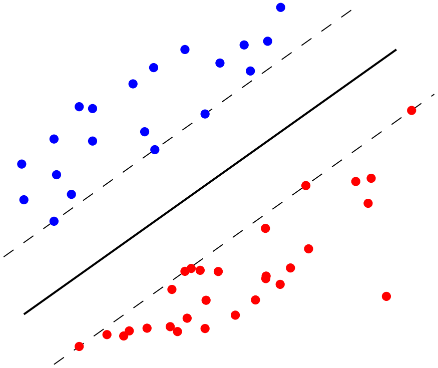
noisy	critical	typical
negative	small positive	large positive

selection cost

- We actually estimate a scaled version of the selection cost (Nicholson, 2002).
- The model for learning should be also used for the estimation.

SVM Confidence Margin

The soft-margin support vector machine (SVM) (Vapnik, 1995) finds a large-confidence hyperplane classifier in the feature space.



- The confidence margin is a meaningful estimate of the intrinsic margin.
- Better than the one used in (Guyon et al., 1996).
- Confidence margin ≤ 1: support vectors critical
- Negative margin noisy

AdaBoost Sample Weight

AdaBoost (Freund & Schapire, 1996) is an algorithm to improve the accuracy of a base learner.

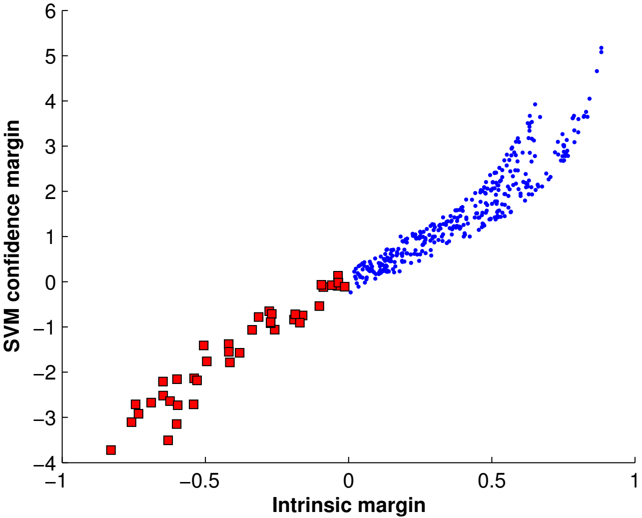
- It iteratively generates an ensemble of base hypotheses.
- It gradually forces the base learner to focus on "hard" examples by giving erroneous examples higher sample weight.

The sample weight is actually a consensus among the base hypotheses on the "hardness" of the example.

- If an example is too "hard", it is probably noisy.
- If an example is too "easy", it is probably typical.
- The negative average sample weight over different iterations is a robust estimate of the intrinsic margin.

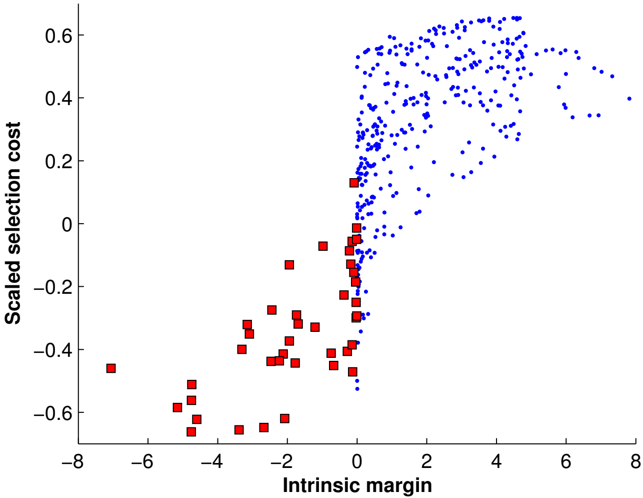
Scatter Plot

3-5-1 NNet



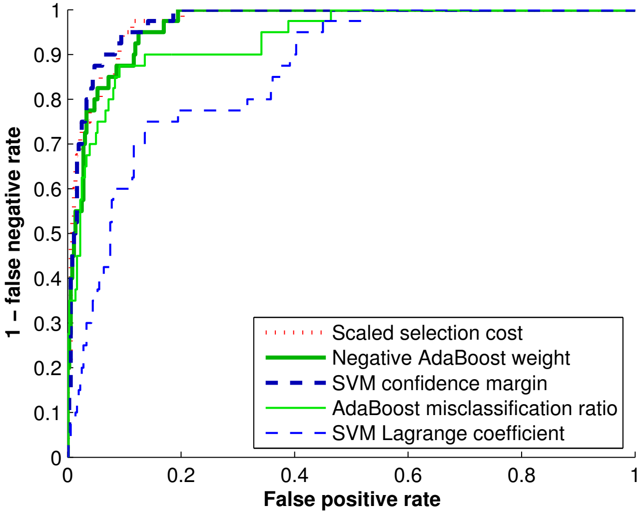
Scatter Plot

Sin (Merler et al., 2004)



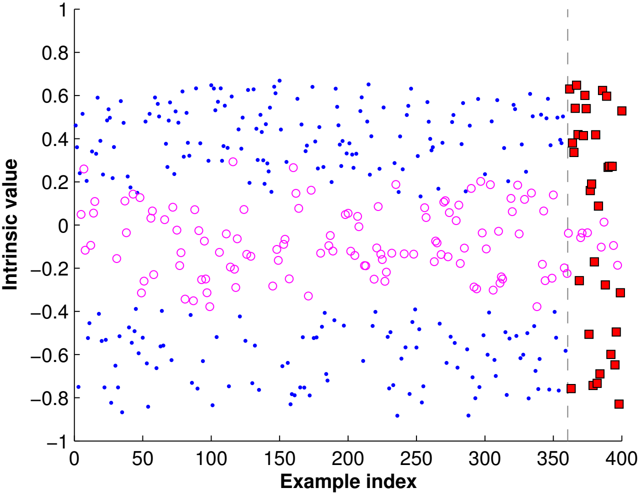
ROC Curves

Sin



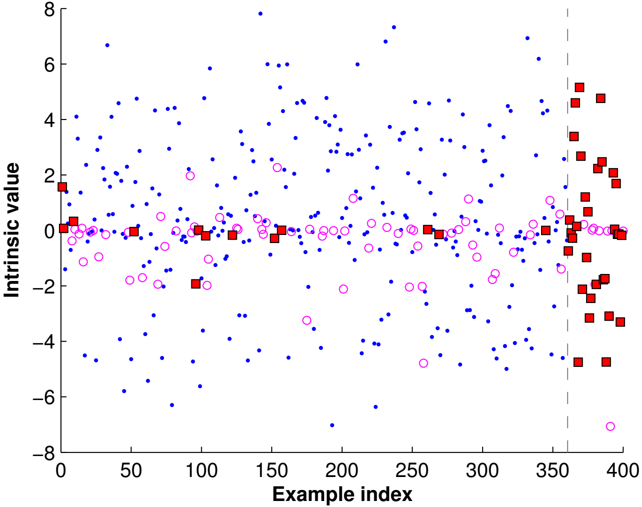
Fingerprint Plot

3-5-1 NNet



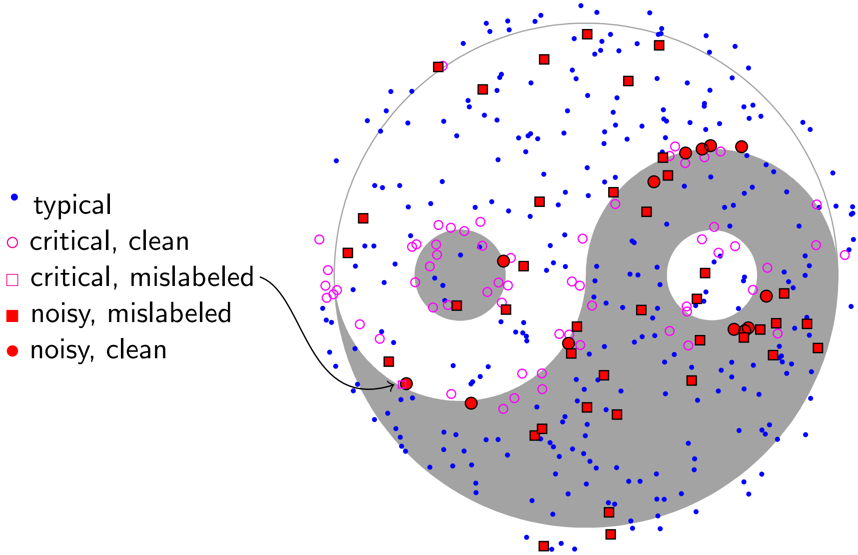
Fingerprint Plot

Sin



2-D Plot

Yin-Yang (http://www.work.caltech.edu/ling/data/yinyang.html)



Real-World Data

Utilize Data Categorization

It is now possible to treat different categories differently.

- Noisy examples: remove
- Critical examples: emphasize
- Typical examples: reduce

dataset	orig. dataset	selection cost	SVM margin	AdaBoost weight
australian	16.65 ± 0.19	15.23 ± 0.20	14.83 ± 0.18	13.92 ± 0.16
breast	4.70 ± 0.11	6.44 ± 0.13	3.40 ± 0.10	3.32 ± 0.10
cleveland	21.64 ± 0.31	18.24 ± 0.30	18.91 ± 0.29	18.56 ± 0.30
german	26.11 ± 0.20	30.12 ± 0.15	24.59 ± 0.20	24.68 ± 0.22
heart	21.93 ± 0.43	17.33 ± 0.34	17.59 ± 0.32	18.52 ± 0.37
pima	26.14 ± 0.20	35.16 ± 0.20	24.02 ± 0.19	25.15 ± 0.20
votes84	5.20 ± 0.14	6.45 ± 0.17	5.03 ± 0.13	4.91 ± 0.13

Conclusion

Contributions

1. Proposed 3 methods for automatically categorizing examples.
   - The methods are from different parts of learning theory.
   - They all gave reasonable categorization results.
2. Tested learning with categorized data.
   - A simple strategy is enough to improve learning.
   - The categorization results can be used in conjunction with a large variety of learning algorithms.
3. Showed experimentally data categorization is powerful.

Future Work

- Estimate the optimal thresholds (say, using a validation set)
- Better utilize the categorization in learning
- Extend the framework to problems other than classification