This lecture notes is based on the following paper:
H. He and E. A. Garcia, "Learning from Imbalanced Data," IEEE Trans. Knowledge and Data Engineering, vol. 21, issue 9, pp. 1263-1284, 2009

Learning from Imbalanced Data

Prof. Haibo He

Electrical Engineering

University of Rhode Island, Kingston, RI 02881

Computational Intelligence and Self-Adaptive Systems (CISA) Laboratory

http://www.ele.uri.edu/faculty/he/

Email: he@ele.uri.edu

Learning from Imbalanced Data

1. The problem: Imbalanced Learning
2. The solutions: State-of-the-art
3. The evaluation: Assessment Metrics
4. The future: Opportunities and Challenges

Source: H. He and E. A. Garcia, "Learning from Imbalanced Data," IEEE Trans. Knowledge and Data Engineering, vol. 21, issue 9, pp. 1263-1284, 2009

The Nature of Imbalanced Learning Problem

Source: H. He and E. A. Garcia, "Learning from Imbalanced Data," IEEE Trans. Knowledge and Data Engineering, vol. 21, issue 9, pp. 1263-1284, 2009

The Problem

✓ Explosive availability of raw data
✓ Well-developed algorithms for data analysis

Requirement?

- ***Balanced*** *distribution* of data
- ***Equal*** *costs* of misclassification

What about data in reality?

Source: H. He and E. A. Garcia, "Learning from Imbalanced Data," IEEE Trans. Knowledge and Data Engineering, vol. 21, issue 9, pp. 1263-1284, 2009

Imbalance is Everywhere

- Between-class
- Within-class
- Intrinsic and extrinsic
- Relativity and rarity
- Imbalance and small sample size

Source: H. He and E. A. Garcia, "Learning from Imbalanced Data," IEEE Trans. Knowledge and Data Engineering, vol. 21, issue 9, pp. 1263-1284, 2009

THE UNIVERSITY OF RHODE ISLAND

Growing interest

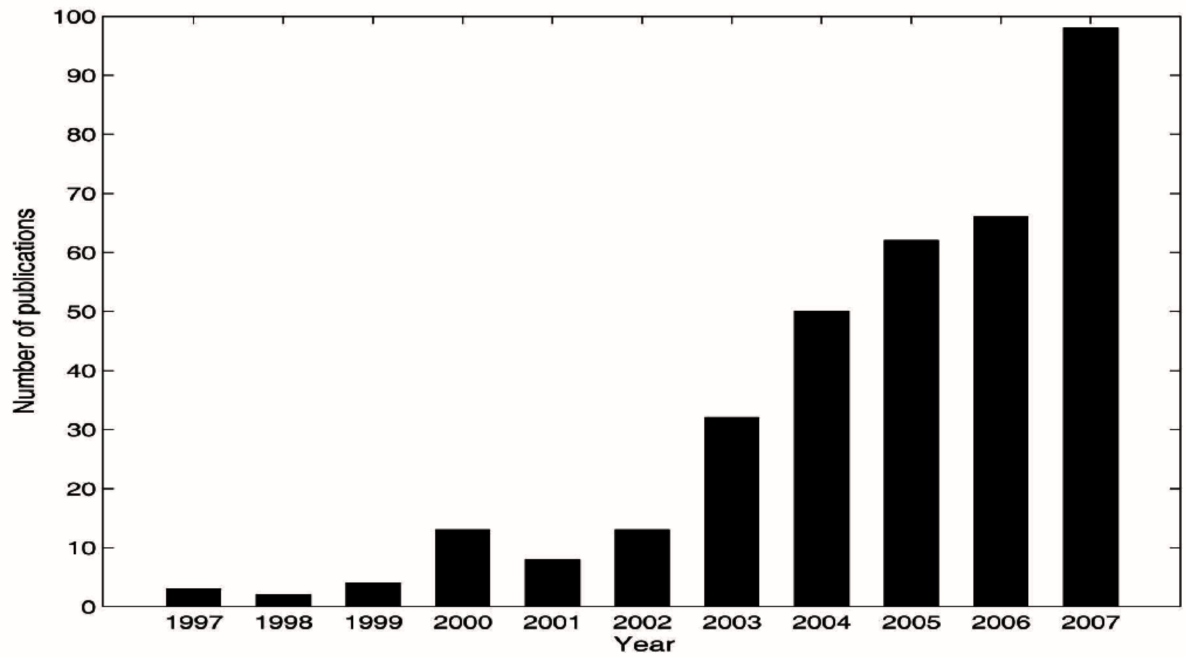


Fig. 1. Number of publications on imbalanced learning.

Source: H. He and E. A. Garcia, "Learning from Imbalanced Data," IEEE Trans. Knowledge and Data Engineering, vol. 21, issue 9, pp. 1263-1284, 2009

The Nature of Imbalance Learning

Mammography Data Set:
An example of ***between-class*** *imbalance*

	Negative/healthy	Positive/cancerous
Number of cases	10,923	260
Category	Majority	Minority
Imbalanced accuracy	≈ 100%	0-10 %

Imbalance can be on the order of 100 : 1 up to 10,000 : 1!

Source: H. He and E. A. Garcia, "Learning from Imbalanced Data," IEEE Trans. Knowledge and Data Engineering, vol. 21, issue 9, pp. 1263-1284, 2009

Intrinsic and ***extrinsic*** imbalance

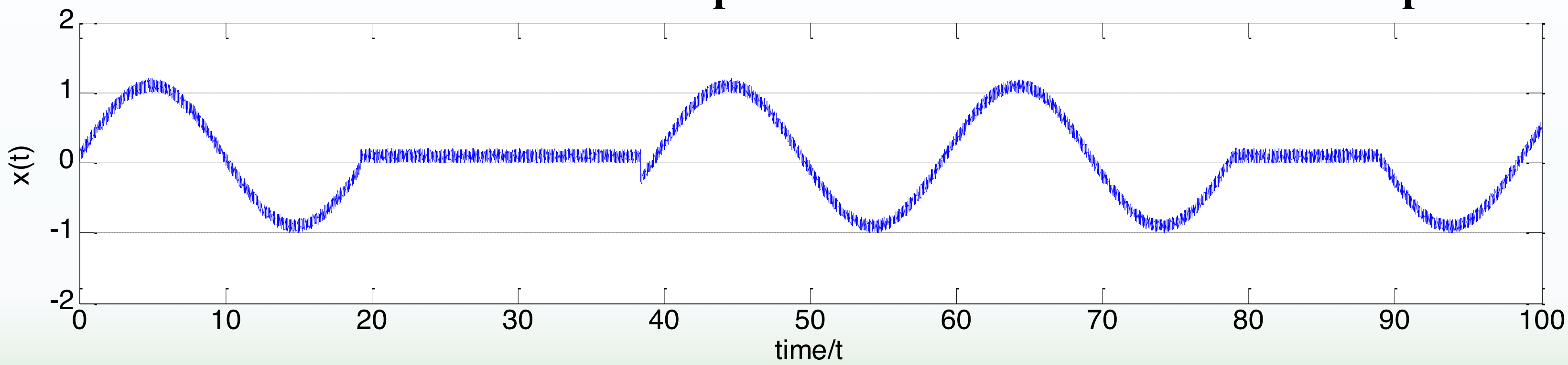
Intrinsic:

- Imbalance due to the nature of the dataspace

Extrinsic:

- Imbalance due to time, storage, and other factors
- **Example:**

Data transmission over a specific interval of time with interruption



Source: H. He and E. A. Garcia, "Learning from Imbalanced Data," IEEE Trans. Knowledge and Data Engineering, vol. 21, issue 9, pp. 1263-1284, 2009

The Nature of Imbalance Learning

Data Complexity

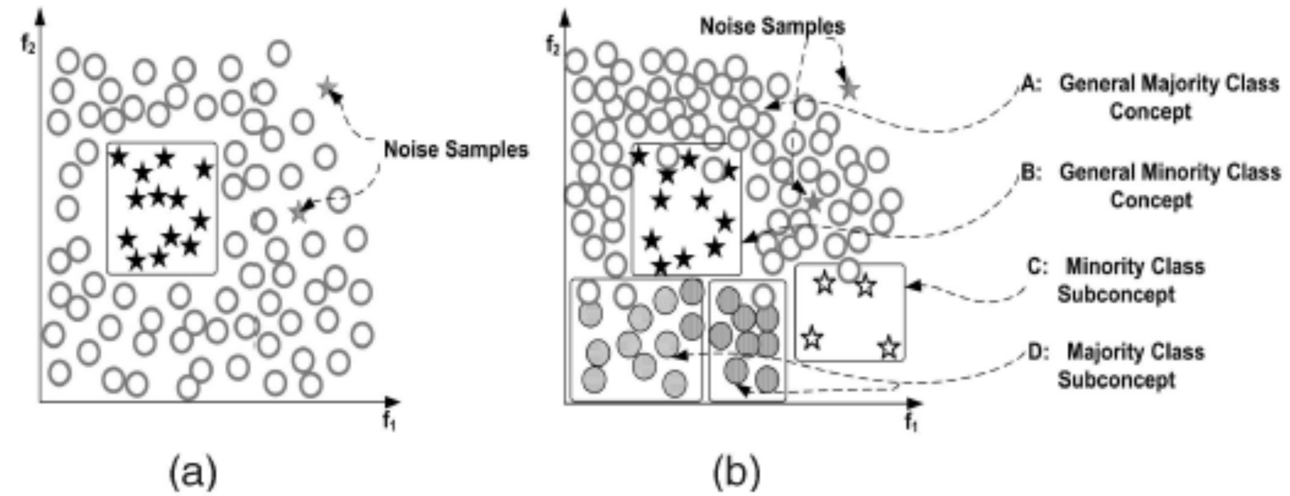


Fig. 2. (a) A data set with a between-class imbalance. (b) A high-complexity data set with both between-class and within-class imbalances, multiple concepts, overlapping, noise, and lack of representative data.

Source: H. He and E. A. Garcia, "Learning from Imbalanced Data," IEEE Trans. Knowledge and Data Engineering, vol. 21, issue 9, pp. 1263-1284, 2009

THE UNIVERSITY OF RHODE ISLAND

Relative imbalance and ***absolute rarity***

Q: $1{,}000{,}000 : 1{,}000 = 1{,}000 : 1$ **?**

- The minority class may be outnumbered, but not necessarily rare
- Therefore they can be accurately learned with little disturbance

Source: H. He and E. A. Garcia, "Learning from Imbalanced Data," IEEE Trans. Knowledge and Data Engineering, vol. 21, issue 9, pp. 1263-1284, 2009

Imbalanced data with ***small sample size***

- Data with high dimensionality and small sample size
  - Face recognition, gene expression

- Challenges with small sample size:
  1. Embedded absolute rarity and within-class imbalances
  2. Failure of generalizing inductive rules by learning algorithms
     - Difficulty in forming good classification decision boundary over *more* features but *less* samples
     - Risk of overfitting

Source: H. He and E. A. Garcia, "Learning from Imbalanced Data," IEEE Trans. Knowledge and Data Engineering, vol. 21, issue 9, pp. 1263-1284, 2009

The Solutions to Imbalanced Learning Problem

Source: H. He and E. A. Garcia, "Learning from Imbalanced Data," IEEE Trans. Knowledge and Data Engineering, vol. 21, issue 9, pp. 1263-1284, 2009

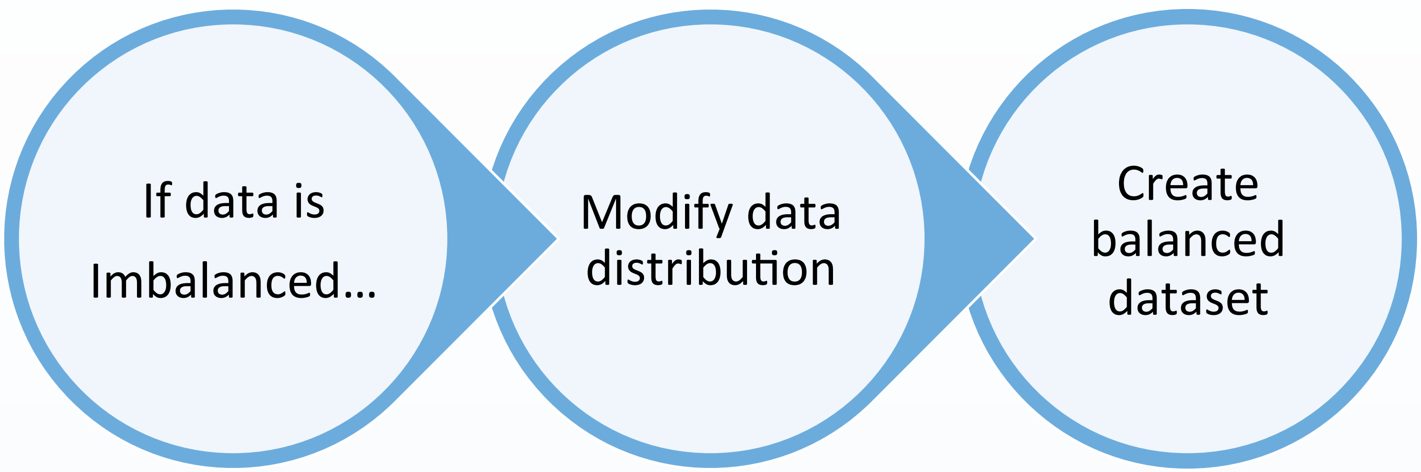
Solutions to imbalanced learning

- Sampling methods
- Cost-sensitive methods
- Kernel and Active Learning methods

Source: H. He and E. A. Garcia, "Learning from Imbalanced Data," IEEE Trans. Knowledge and Data Engineering, vol. 21, issue 9, pp. 1263-1284, 2009

Sampling methods

Sampling methods



Create balance though sampling

Source: H. He and E. A. Garcia, "Learning from Imbalanced Data," IEEE Trans. Knowledge and Data Engineering, vol. 21, issue 9, pp. 1263-1284, 2009

THE UNIVERSITY OF RHODE ISLAND

Sampling methods

Random Sampling

S: training data set; $\boldsymbol{S_{min}}$: set of minority class samples, $\boldsymbol{S_{maj}}$: set of majority class samples; ***E***: generated samples

Random oversampling	*Random undersampling*
• Expand the minority • $\lvert S'_{min}\rvert \leftarrow \lvert S_{min}\rvert + \lvert E\rvert$ • $\lvert S'\rvert \leftarrow \lvert S_{min}\rvert + \lvert S_{maj}\rvert + \lvert E\rvert$ • Overfitting due to multiple “tied” instances	• Shrink the majority • $\lvert S'_{maj}\rvert \leftarrow \lvert S_{maj}\rvert - \lvert E\rvert$ • $\lvert S'\rvert \leftarrow \lvert S_{min}\rvert + \lvert S_{maj}\rvert - \lvert E\rvert$ • Loss of important concepts

Source: H. He and E. A. Garcia, “Learning from Imbalanced Data,” IEEE Trans. Knowledge and Data Engineering, vol. 21, issue 9, pp. 1263-1284, 2009

Sampling methods

Informed Undersampling

- *EasyEnsemble*
  - ***Unsupervised****: use random subsets of the majority class to create balance and form multiple classifiers*
- *BalanceCascade*
  - ***Supervised****: iteratively create balance and pull out redundant samples in majority class to form a final classifier*

1. Generate $E \subset S_{maj}$ $(s.t. |E| = |S_{\min}|)$, and $N = \{E \cup S_{\min}\}$
2. Induce $H(n)$
3. Identify N_{maj}^{*} as samples from N that are correctly classified
4. Remove N_{maj}^{*} from S_{maj}
5. Repeat (1) and induce $H(n+1)$ until stopping criteria is met

Source: H. He and E. A. Garcia, "Learning from Imbalanced Data," IEEE Trans. Knowledge and Data Engineering, vol. 21, issue 9, pp. 1263-1284, 2009

Informed Undersampling

- ***Undersampling using K-nearest neighbor (KNN) classifier***
  - NearMiss-1, NearMiss-2, NearMiss-3, and the "most distant" method
  - NearMiss-2 provides competitive results for imbalanced learning
- ***One-sided selection (OSS)***
  - Selects representative subset E from the majority class
  - Combine with the minority class $N = \{E \cup S_{\text{min}}\}$
  - Refine N with data cleaning techniques

Source: H. He and E. A. Garcia, "Learning from Imbalanced Data," IEEE Trans. Knowledge and Data Engineering, vol. 21, issue 9, pp. 1263-1284, 2009

Sampling methods

Synthetic Sampling with Data Generation

- Synthetic minority oversampling technique (SMOTE)
  - Creates artificial minority class data using feature space similarities
  - For $\forall x_i \in S_{\min}$
    1. Randomly choose one of the k nearest neighbor $\hat{x}_i$;
    2. Create a new sample $x_{new} = x_i + (\hat{x}_i - x_i) \times \delta$, where δ is a uniformly distributed random variable.

Source: H. He and E. A. Garcia, "Learning from Imbalanced Data," IEEE Trans. Knowledge and Data Engineering, vol. 21, issue 9, pp. 1263-1284, 2009

Sampling methods

Synthetic Sampling with Data Generation

- Synthetic minority oversampling technique (SMOTE)

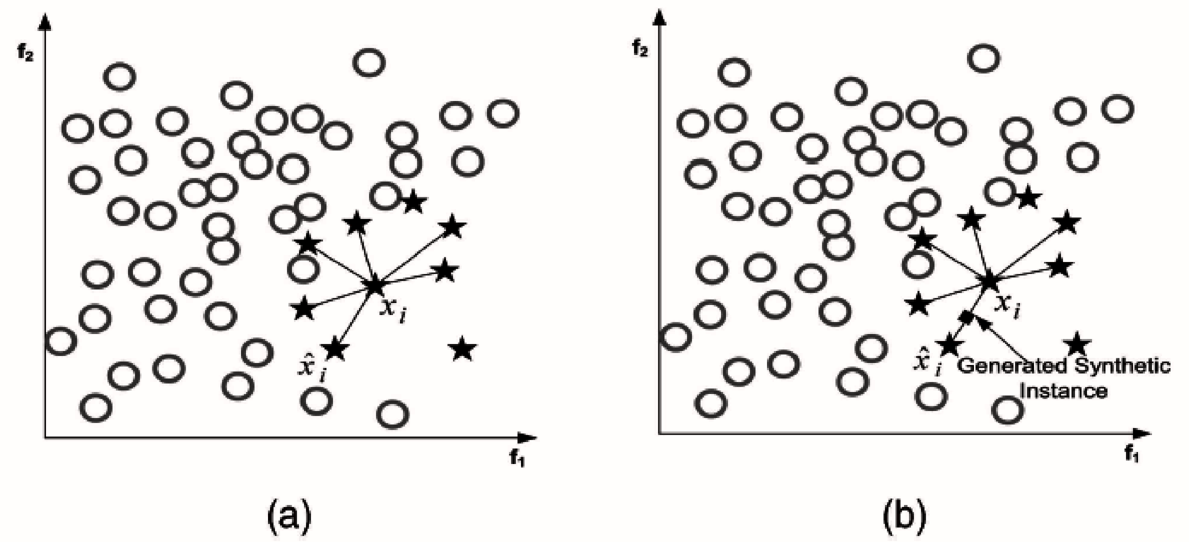


Fig. 3. (a) Example of the K-nearest neighbors for the x_i example under consideration ($K = 6$). (b) Data creation based on euclidian distance.

Source: H. He and E. A. Garcia, "Learning from Imbalanced Data," IEEE Trans. Knowledge and Data Engineering, vol. 21, issue 9, pp. 1263-1284, 2009

Adaptive Synthetic Sampling

- Overcomes over generalization in SMOTE algorithm
  - Border-line-SMOTE:
    1. Determine the set of m-nearest neighbors for each $x_i \in S_{min}$, call it $S_{i:m-NN}$
    2. Identify the number of nearest neighbors in majority class, i.e., $|S_{i:m-NN} \cap S_{maj}|$
    3. Select x_i that satisfies: $\frac{m}{2} \leq |S_{i:m-NN} \cap S_{maj}| < m$

Source: H. He and E. A. Garcia, "Learning from Imbalanced Data," IEEE Trans. Knowledge and Data Engineering, vol. 21, issue 9, pp. 1263-1284, 2009

Sampling methods

Adaptive Synthetic Sampling

- Overcomes over generalization in SMOTE algorithm
  - Border-line-SMOTE

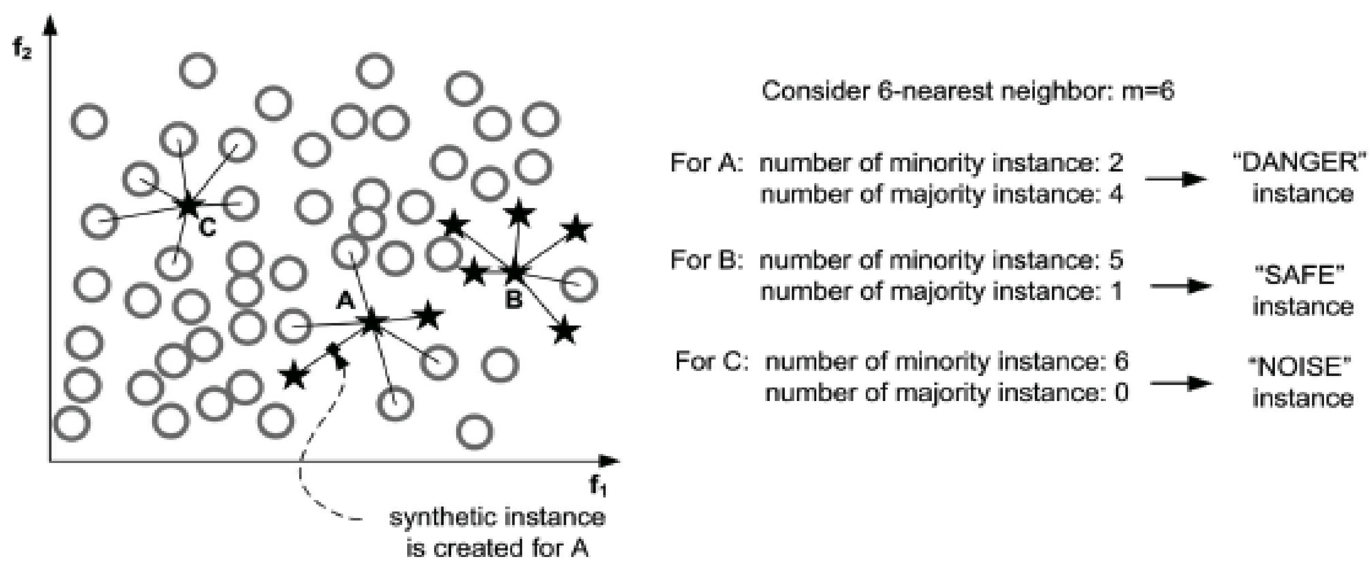


Fig. 4. Data creation based on Borderline instance.

Source: H. He and E. A. Garcia, "Learning from Imbalanced Data," IEEE Trans. Knowledge and Data Engineering, vol. 21, issue 9, pp. 1263-1284, 2009

Sampling methods

Adaptive Synthetic Sampling

- Overcomes over generalization in SMOTE algorithm
  - ADASYN

1. Calculate number of synthetic samples $G = \left(|S_{maj}| - |S_{\min}|\right) \times \beta$
2. for each $x_i \in S_{\min}$, find k-nearest neighbors and calculate ratio $\Gamma_i = \frac{\Delta_i / K}{Z}, i = 1, \ldots, |S_{min}|$ as a distribution function;
3. Identify the number of synthetic samples to be generated for x_i by $g_i = \Gamma_i \times G$
4. Generate x_{new} using SMOTE algorithm: $x_{new} = x_i + (\hat{x}_i - x_i) \times \delta$

Source: H. He and E. A. Garcia, "Learning from Imbalanced Data," IEEE Trans. Knowledge and Data Engineering, vol. 21, issue 9, pp. 1263-1284, 2009

Sampling with Data Cleaning

- Tomek links
  1. Given a pair (x_i, x_j) where $x_i \in S_{\min}$, $x_j \in S_{maj}$, and the distance between them as $d(x_i, x_j)$
  2. If there is no instance x_k $s.t. d(x_i, x_k) <. d(x_i, x_j)$ or $d(x_j, x_k) < d(x_i, x_j)$, then (x_i, x_j) is called a Tomek link
- Clean up unwanted inter-class overlapping after synthetic sampling
- Examples:
  - OSS, condensed nearest neighbor and Tomek links (CNN + Tomek links), neighborhood cleaning rule (NCL) based on edited nearest neighbor (ENN), SMOTE+ENN, and SMOTE+Tomek

Source: H. He and E. A. Garcia, "Learning from Imbalanced Data," IEEE Trans. Knowledge and Data Engineering, vol. 21, issue 9, pp. 1263-1284, 2009

Sampling methods

Sampling with Data Cleaning

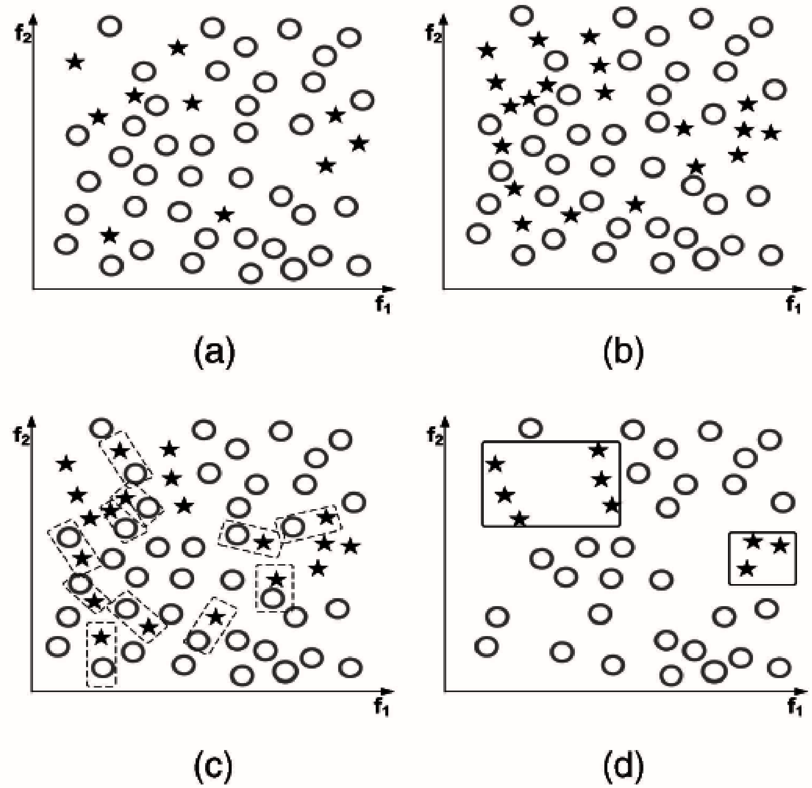


Fig. 5. (a) Original data set distribution. (b) Post-SMOTE data set. (c) The identified Tomek Links. (d) The data set after removing Tomek links.

Source: H. He and E. A. Garcia, "Learning from Imbalanced Data," IEEE Trans. Knowledge and Data Engineering, vol. 21, issue 9, pp. 1263-1284, 2009

Cluster-based oversampling (CBO) method

1. For the majority class S_{maj} with m_{maj} clusters
   I. Oversample each cluster $C_{maj:j} \subset S_{maj}, j = 1, \dots, m_{maj}$ except the largest $C_{maj:\max}$, so that for $\forall j, |C_{maj:j}| = |C_{maj:\max}|$
   II. Calculate the number of majority class examples after oversampling as N_{CBO}
2. For the minority class S_{min} with m_{min} clusters
   I. Oversample each cluster $C_{min:i} \subset S_{min}, i = 1, \dots, m_{\min}$ to be of the same size $N_{CBO}/m_{\min}$, so that for $\forall i, |C_{min:i}| = N_{CBO}/m_{\min}$

Source: H. He and E. A. Garcia, "Learning from Imbalanced Data," IEEE Trans. Knowledge and Data Engineering, vol. 21, issue 9, pp. 1263-1284, 2009

Sampling methods

CBO Method

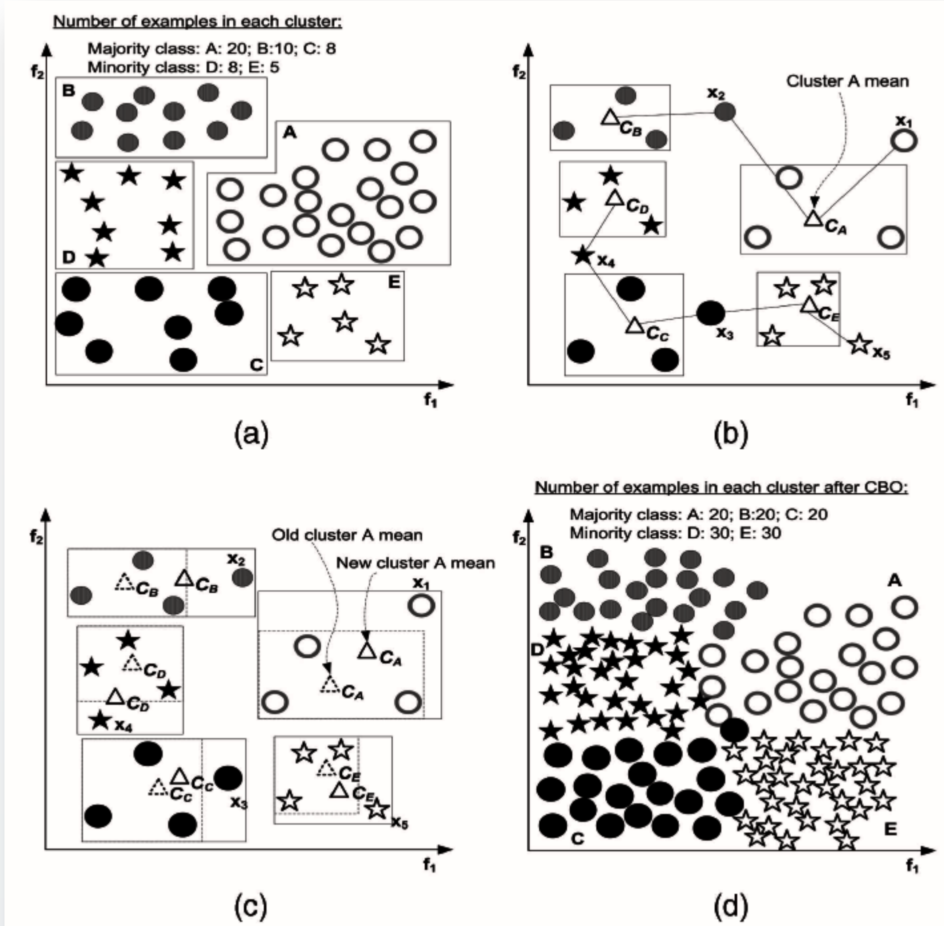


Fig. 6. (a) Original data set distribution. (b) Distance vectors of examples and cluster means. (c) Newly defined cluster means and cluster borders. (d) The data set after cluster-based oversampling method.

Source: H. He and E. A. Garcia, "Learning from Imbalanced Data," IEEE Trans. Knowledge and Data Engineering, vol. 21, issue 9, pp. 1263-1284, 2009

Integration of Sampling and Boosting

1. SMOTEBoost
   - SMOTE + Adaboost.M2
   - Introduces synthetic sampling at each boosting iteration
2. DataBoost-IM
   - AdaBoost.M1
   - Generate synthetic date of hard-to-learn samples for both majority and minority classes (usually $|E_{maj}| < |E_{min}|$)
3. JOUS-Boost

Source: H. He and E. A. Garcia, "Learning from Imbalanced Data," IEEE Trans. Knowledge and Data Engineering, vol. 21, issue 9, pp. 1263-1284, 2009

Integration of Sampling and Boosting

DataBoost-IM

1. Collect the set E of top misclassified samples (hard-to-learn samples) for both classes with subsets $E_{maj} \subset E$ and $E_{min} \subset E$
2. Identify M_L seeds from E_{maj} and M_S seeds from E_{min}, where

   $M_L = \min(\frac{|S_{maj}|}{|S_{min}|}, |E_{maj}|)$ and $M_S = \min(\frac{|S_{maj}| \times M_L}{|S_{\min}|}, |E_{\min}|)$
3. Generate synthetic set E_{syn} with subsets for both classes:

   $E_{s\min} \subset E_{syn}$ and $E_{smaj} \subset E_{syn}$

   $s.t. |E_{smin}| = M_S \times |S_{\min}|$ and $\left|E_{smaj}\right| = M_L \times |S_{maj}|$

Source: H. He and E. A. Garcia, "Learning from Imbalanced Data," IEEE Trans. Knowledge and Data Engineering, vol. 21, issue 9, pp. 1263-1284, 2009

Cost-Sensitive Methods

Source: H. He and E. A. Garcia, "Learning from Imbalanced Data," IEEE Trans. Knowledge and Data Engineering, vol. 21, issue 9, pp. 1263-1284, 2009

Cost-Sensitive Learning Framework

- Define the cost of misclassifying a majority to a minority as $C(Min, Maj)$
- Typically $C(Maj, Min) > C(Min, Maj)$
- Minimize the overall cost - usually the *Bayes conditional risk* - on the training data set

$$R(i|x) = \sum_j P(j|x)C(i,j)$$

True Class j

Predicted Class i	1	2	...	k
1	$C(1,1)$	$C(1,2)$	...	$C(1,k)$
2	$C(2,1)$	...	...	$\vdots$
$\vdots$	$\vdots$	...	...	$\vdots$
k	$C(k,1)$	...	...	$C(k,k)$

Fig. 7. Multiclass cost matrix.

Source: H. He and E. A. Garcia, "Learning from Imbalanced Data," IEEE Trans. Knowledge and Data Engineering, vol. 21, issue 9, pp. 1263-1284, 2009

Cost-Sensitive methods

Cost-Sensitive Dataspace Weighting with Adaptive Boosting

- Iteratively update the distribution function D_t of the training data according to error of current hypothesis h_t and cost factor C_i
  - Weight updating parameter $\alpha_t = \frac{1}{2}\ln(\frac{1-\varepsilon_t}{\varepsilon_t})$
  - Error of hypothesis h_t: $\varepsilon_t = \sum_{i:h_t(x_i)\neq y_i} D_t(i)$

Source: H. He and E. A. Garcia, "Learning from Imbalanced Data," IEEE Trans. Knowledge and Data Engineering, vol. 21, issue 9, pp. 1263-1284, 2009

Cost-Sensitive methods

Cost-Sensitive Dataspace Weighting with Adaptive Boosting

- Given D_t, h_t, C_i, α_t, and ε_t

1. AdaC1: $D_{t+1}(i) = \frac{D_t(i)\, exp(-\alpha_t C_i h_t(x_i) y_i)}{Z_t}$

2. AdaC2: $D_{t+1}(i) = \frac{C_i D_t(i)\, exp(-\alpha_t h_t(x_i) y_i)}{Z_t}$

3. AdaC3: $D_{t+1}(i) = \frac{C_i D_t(i)\, exp(-\alpha_t C_i h_t(x_i) y_i)}{Z_t}$

4. AdaCost: $D_{t+1}(i) = \frac{D_t(i)\, exp(-\alpha_t h_t(x_i) y_i \beta_i)}{Z_t}$,

   $\beta_i = \beta(sign(y_i, h_t(x_i)), C_i)$

Source: H. He and E. A. Garcia, "Learning from Imbalanced Data," IEEE Trans. Knowledge and Data Engineering, vol. 21, issue 9, pp. 1263-1284, 2009

Cost-Sensitive Decision Trees

1. Cost-sensitive adjustments for the decision threshold
   - The final decision threshold shall yield the most dominant point on the ROC curve
2. Cost-sensitive considerations for split criteria
   - The impurity function shall be insensitive to unequal costs
3. Cost-sensitive pruning schemes
   - The probability estimate at each node needs improvement to reduce removal of leaves describing the minority concept
   - Laplace smoothing method and Laplace pruning techniques

Source: H. He and E. A. Garcia, "Learning from Imbalanced Data," IEEE Trans. Knowledge and Data Engineering, vol. 21, issue 9, pp. 1263-1284, 2009

Cost-Sensitive Neural Network

Four ways of applying cost sensitivity in neural networks

Modifying probability estimate of outputs	Altering outputs directly	Modify learning rate	Replacing error-minimizing function
• *Applied only at testing stage* • *Maintain original neural networks*	• *Bias neural networks during training to focus on expensive class*	• *Set η higher for costly examples and lower for low-cost examples*	• *Use expected cost minimization function instead*

Source: H. He and E. A. Garcia, "Learning from Imbalanced Data," IEEE Trans. Knowledge and Data Engineering, vol. 21, issue 9, pp. 1263-1284, 2009

Kernel-based learning framework

- Based on statistical learning and Vapnik-Chervonenkis (VC) dimensions
- Problems with Kernel-based support vector machines (SVMs)
    1. Support vectors from the minority concept may contribute less to the final hypothesis
    2. Optimal hyperplane is also biased toward the majority class

To minimize the **total** error → Biased toward the majority

Source: H. He and E. A. Garcia, "Learning from Imbalanced Data," IEEE Trans. Knowledge and Data Engineering, vol. 21, issue 9, pp. 1263-1284, 2009

Integration of Kernel Methods with Sampling Methods

1. SMOTE with Different Costs (SDCs) method
2. Ensembles of over/under-sampled SVMs
3. SVM with asymmetric misclassification cost
4. Granular Support Vector Machines—Repetitive Undersampling (GSVM-RU) algorithm

Source: H. He and E. A. Garcia, "Learning from Imbalanced Data," IEEE Trans. Knowledge and Data Engineering, vol. 21, issue 9, pp. 1263-1284, 2009

Kernel Modification Methods

1. Kernel classifier construction
   - Orthogonal forward selection (OFS) and Regularized orthogonal weighted least squares (ROWLSs) estimator
2. SVM class boundary adjustment
   - Boundary movement (BM), biased penalties (BP), class-boundary alignment(CBA), kernel-boundary alignment (KBA)
3. Integrated approach
   - Total margin-based adaptive fuzzy SVM (TAF-SVM)
4. K-category proximal SVM (PSVM) with Newton refinement
5. Support cluster machines (SCMs), Kernel neural gas (KNG), P2PKNNC algorithm, hybrid kernel machine ensemble (HKME) algorithm, Adaboost relevance vector machine (RVM), …

Source: H. He and E. A. Garcia, "Learning from Imbalanced Data," IEEE Trans. Knowledge and Data Engineering, vol. 21, issue 9, pp. 1263-1284, 2009

Active Learning Methods

- SVM-based active learning

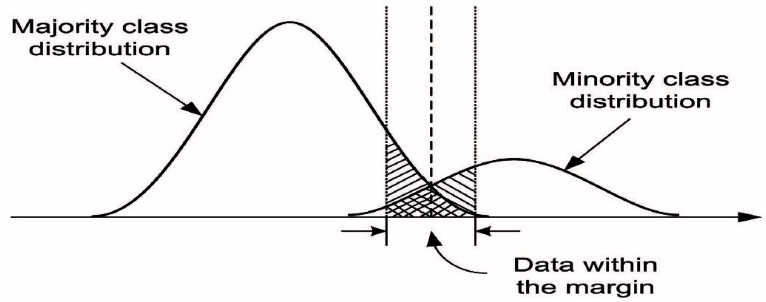


Fig. 8. Data imbalance ratio within and outside the margin [98].

- Active learning with sampling techniques
  - Undersampling and oversampling with active learning for the word sense disambiguation (WSD) imbalanced learning
  - New stopping mechanisms based on maximum confidence and minimal error
  - Simple active learning heuristic (SALH) approach

Source: H. He and E. A. Garcia, "Learning from Imbalanced Data," IEEE Trans. Knowledge and Data Engineering, vol. 21, issue 9, pp. 1263-1284, 2009

Additional methods

One-class learning/novelty detection methods

Mahalanobis-Taguchi System

Rank metrics and multitask learning

- Combination of imbalanced data and the small sample size problem

Multiclass imbalanced learning

- AdaC2.M1
- Rescaling approach for multiclass cost-sensitive neural networks
- the ensemble knowledge for imbalance sample sets (eKISS) method

Source: H. He and E. A. Garcia, "Learning from Imbalanced Data," IEEE Trans. Knowledge and Data Engineering, vol. 21, issue 9, pp. 1263-1284, 2009

The Evaluation of Imbalanced Learning Problem

THE UNIVERSITY OF RHODE ISLAND

Assessment Metrics

How to evaluate the performance of imbalanced learning algorithms ?

1. Singular assessment metrics
2. Receiver operating characteristics (ROC) curves
3. Precision-Recall (PR) Curves
4. Cost Curves
5. Assessment Metrics for Multiclass Imbalanced Learning

Source: H. He and E. A. Garcia, "Learning from Imbalanced Data," IEEE Trans. Knowledge and Data Engineering, vol. 21, issue 9, pp. 1263-1284, 2009

Singular assessment metrics

		True class: p	True class: n
Hypothesis output	Y	TP (True Positives)	FP (False Positives)
	N	FN (False Negatives)	TN (True Negatives)
Column counts:		P_C	N_C

Fig. 9. Confusion matrix for performance evaluation.

$$Accuracy = \frac{TP + TN}{P_C + N_C}$$

$$ErrorRate = 1 - accuracy$$

- Limitations of accuracy – sensitivity to data distributions

Assessment Metrics

Singular assessment metrics

$$Precision = \frac{TP}{TP + FP},$$

$$Recall = \frac{TP}{TP + FN},$$

- Insensitive to data distributions

Source: H. He and E. A. Garcia, "Learning from Imbalanced Data," IEEE Trans. Knowledge and Data Engineering, vol. 21, issue 9, pp. 1263-1284, 2009

Assessment Metrics

Singular assessment metrics

More comprehensive metrics

$$F\text{-}Measure = \frac{(1+\beta)^2 \cdot Recall \cdot Precision}{\beta^2 \cdot Recall + Precision}$$

$\beta = 1$, usually

$$G\text{-}mean = \sqrt{\frac{TP}{TP+FN} \times \frac{TN}{TN+FP}}$$

Source: H. He and E. A. Garcia, "Learning from Imbalanced Data," IEEE Trans. Knowledge and Data Engineering, vol. 21, issue 9, pp. 1263-1284, 2009

Assessment Metrics

Receive Operating Characteristics (ROC) curves

- $TP_{rate} = \frac{TP}{P_C}$
- $FP_{rate} = \frac{FP}{N_C}$
- Area under the curve (AUC)

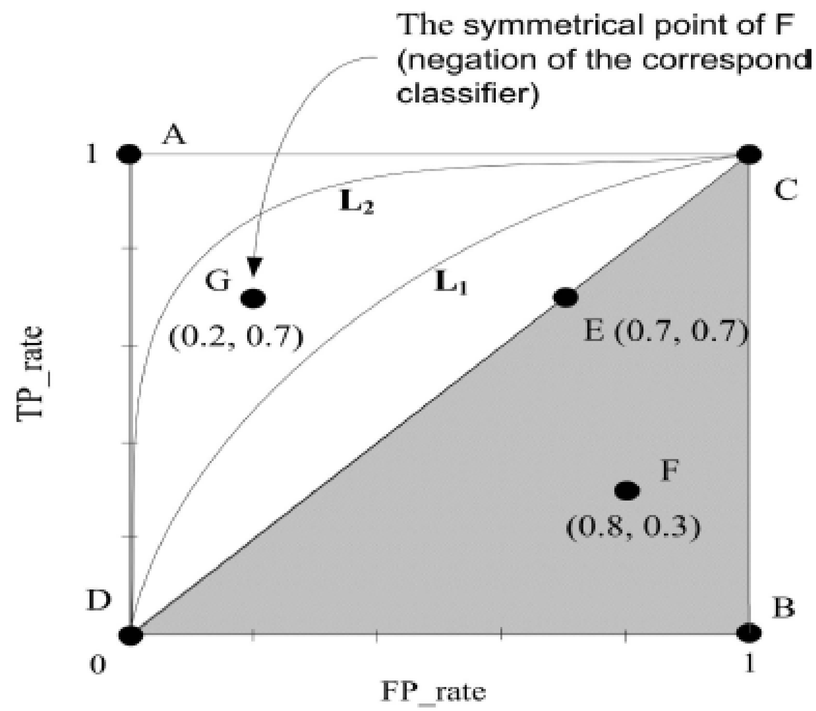


Fig. 10. ROC curve representation.

Source: H. He and E. A. Garcia, "Learning from Imbalanced Data," IEEE Trans. Knowledge and Data Engineering, vol. 21, issue 9, pp. 1263-1284, 2009

Precision-Recall (PR) curves

- Plotting the precision rate over the recall rate
- A curve dominates in ROC space (resides in the upper-left hand) ***if and only if*** it dominates (resides in the upper-right hand) in PR space
- PR space has all the analogous benefits of ROC space
- Provide more informative representations of performance assessment under highly imbalanced data

Source: H. He and E. A. Garcia, "Learning from Imbalanced Data," IEEE Trans. Knowledge and Data Engineering, vol. 21, issue 9, pp. 1263-1284, 2009

Assessment Metrics

Cost Curves

- $PCF(+)$: the probability of an example being from the positive class
- Expected cost: $E[C] = (1 - TP - FP) \times PCF(+) + FP$

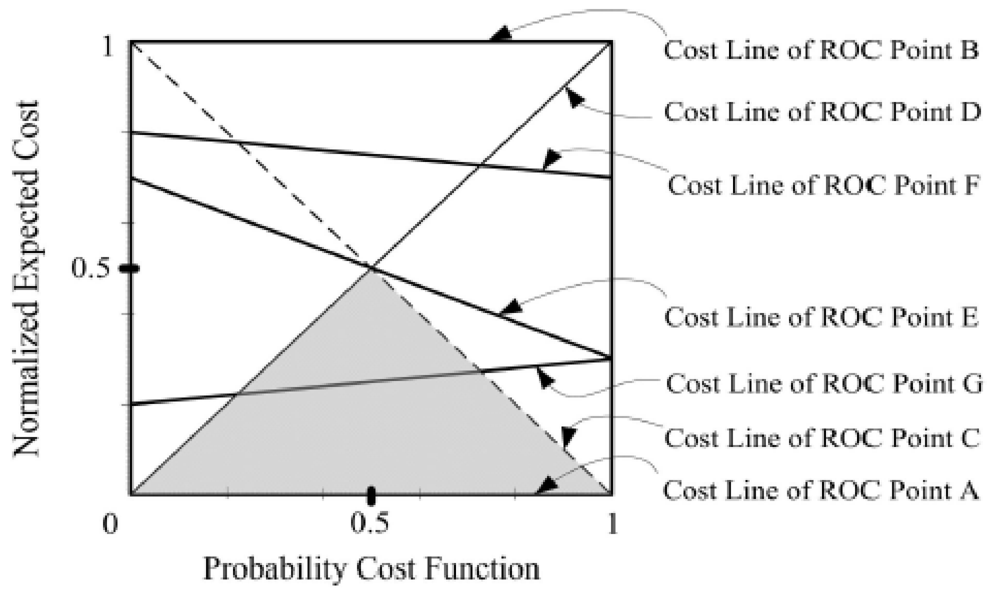


Fig. 11. Cost curve representation.

Source: H. He and E. A. Garcia, "Learning from Imbalanced Data," IEEE Trans. Knowledge and Data Engineering, vol. 21, issue 9, pp. 1263-1284, 2009

The Future of Imbalanced Learning Problem

Source: H. He and E. A. Garcia, "Learning from Imbalanced Data," IEEE Trans. Knowledge and Data Engineering, vol. 21, issue 9, pp. 1263-1284, 2009

Opportunities and Challenges

Understanding the Fundamental Problem

Need of a Uniform Benchmark Platform

Need of Standardized Evaluation Practices

Semi-supervised Learning from Imbalanced data

Source: H. He and E. A. Garcia, "Learning from Imbalanced Data," IEEE Trans. Knowledge and Data Engineering, vol. 21, issue 9, pp. 1263-1284, 2009

Opportunities and Challenges

Understanding the Fundamental Problem

1. What kind of assumptions will make imbalanced learning algorithms work better compared to learning from the original distributions?
2. To what degree should one balance the original data set?
3. How do imbalanced data distributions affect the computational complexity of learning algorithms?
4. What is the general error bound given an imbalanced data distribution?

Source: H. He and E. A. Garcia, "Learning from Imbalanced Data," IEEE Trans. Knowledge and Data Engineering, vol. 21, issue 9, pp. 1263-1284, 2009

THE UNIVERSITY OF RHODE ISLAND

Opportunities and Challenges

Need of a Uniform Benchmark Platform

1. Lack of a uniform benchmark for standardized performance assessments
2. Lack of data sharing and data interoperability across different disciplinary domains;
3. Increased procurement costs, such as time and labor, for the research community as a whole group since each research group is required to collect and prepare their own data sets.

Source: H. He and E. A. Garcia, "Learning from Imbalanced Data," IEEE Trans. Knowledge and Data Engineering, vol. 21, issue 9, pp. 1263-1284, 2009

THE UNIVERSITY OF RHODE ISLAND

Opportunities and Challenges

Need of Standardized Evaluation Practices

- Establish the practice of using the curve-based evaluation techniques
- A standardized set of evaluation practices for proper comparisons

Source: H. He and E. A. Garcia, "Learning from Imbalanced Data," IEEE Trans. Knowledge and Data Engineering, vol. 21, issue 9, pp. 1263-1284, 2009

Opportunities and Challenges

Incremental Learning from Imbalanced Data Streams

1. How can we autonomously adjust the learning algorithm if an imbalance is introduced in the middle of the learning period?
2. Should we consider rebalancing the data set during the incremental learning period? If so, how can we accomplish this?
3. How can we accumulate previous experience and use this knowledge to adaptively improve learning from new data?
4. How do we handle the situation when newly introduced concepts are also imbalanced (i.e., the imbalanced concept drifting issue)?

Source: H. He and E. A. Garcia, "Learning from Imbalanced Data," IEEE Trans. Knowledge and Data Engineering, vol. 21, issue 9, pp. 1263-1284, 2009

Opportunities and Challenges

Semi-supervised Learning from Imbalanced Data

1. How can we identify whether an unlabeled data example came from a balanced or imbalanced underlying distribution?
2. Given an imbalanced training data with labels, what are the effective and efficient methods for recovering the unlabeled data examples?
3. What kind of biases may be introduced in the recovery process (through the conventional semi-supervised learning techniques) given imbalanced, labeled data?

Source: H. He and E. A. Garcia, "Learning from Imbalanced Data," IEEE Trans. Knowledge and Data Engineering, vol. 21, issue 9, pp. 1263-1284, 2009

Reference:

This lecture notes is based on the following paper:

H. He and E. A. Garcia, "Learning from Imbalanced Data," IEEE Trans. Knowledge and Data Engineering, vol. 21, issue 9, pp. 1263-1284, 2009

Should you have any comments or suggestions regarding this lecture note, please feel free to contact Dr. Haibo He at he@ele.uri.edu

Web: http://www.ele.uri.edu/faculty/he/