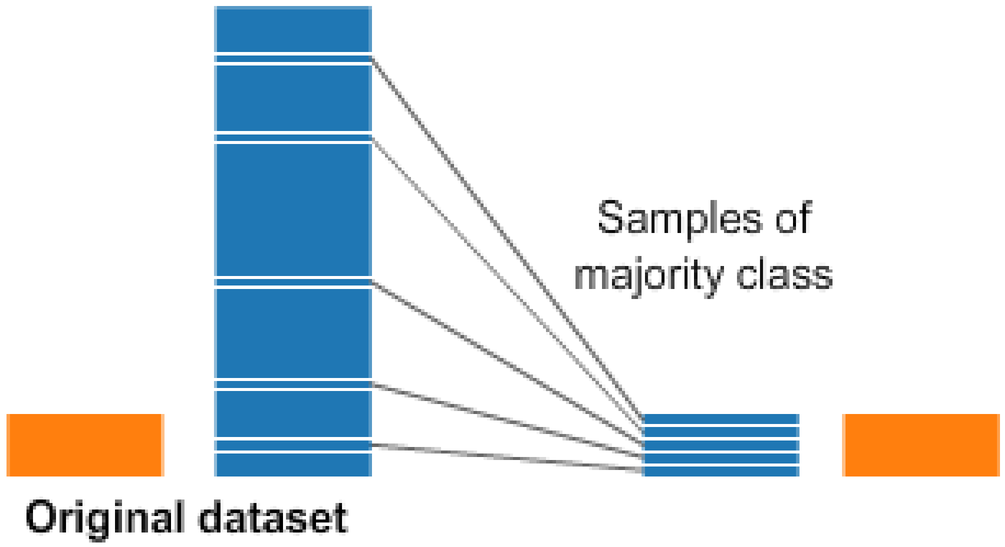
Handling imbalanced datasets

Christian Salvatore
Scuola Universitaria Superiore IUSS Pavia

christian.salvatore@iusspavia.it

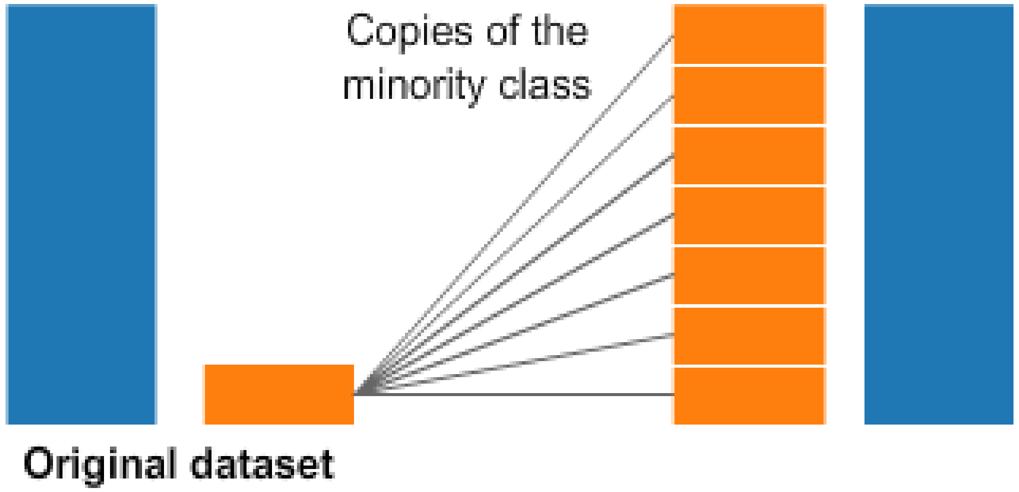
Handling imbalanced datasets

1. Statistical undersampling of the dataset



Handling imbalanced datasets

2. Statistical oversampling of the dataset



3. Synthetic oversampling of the dataset

Synthetic Minority Oversampling Technique (SMOTE)

“The minority class is over-sampled by taking each minority class sample and introducing synthetic examples along the line segments joining any/all of the k minority class nearest neighbors. Depending upon the amount of over-sampling required, neighbors from the k nearest neighbors are randomly chosen

Journal of Artificial Intelligence Research 16 (2002) 321–357 Submitted 09/01; published 06/02

SMOTE: Synthetic Minority Over-sampling Technique

Nitesh V. Chawla CHAWLA@CSEE.USF.EDU
Department of Computer Science and Engineering, ENB 118
University of South Florida
4202 E. Fowler Ave.
Tampa, FL 33620-5399, USA

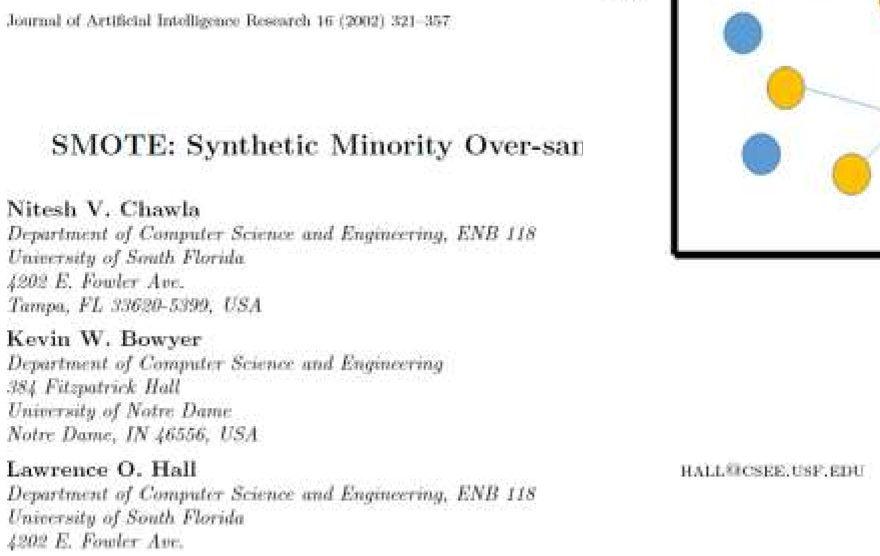
Kevin W. Bowyer KWB@CSE.ND.EDU
Department of Computer Science and Engineering
384 Fitzpatrick Hall
University of Notre Dame
Notre Dame, IN 46556, USA

Lawrence O. Hall HALL@CSEE.USF.EDU
Department of Computer Science and Engineering, ENB 118
University of South Florida
4202 E. Fowler Ave.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, *16*, 321-357.

3. Synthetic oversampling of the dataset

Synthetic Minority Oversampling Technique (SMOTE)


Journal of Artificial Intelligence Research 16 (2002) 321–357

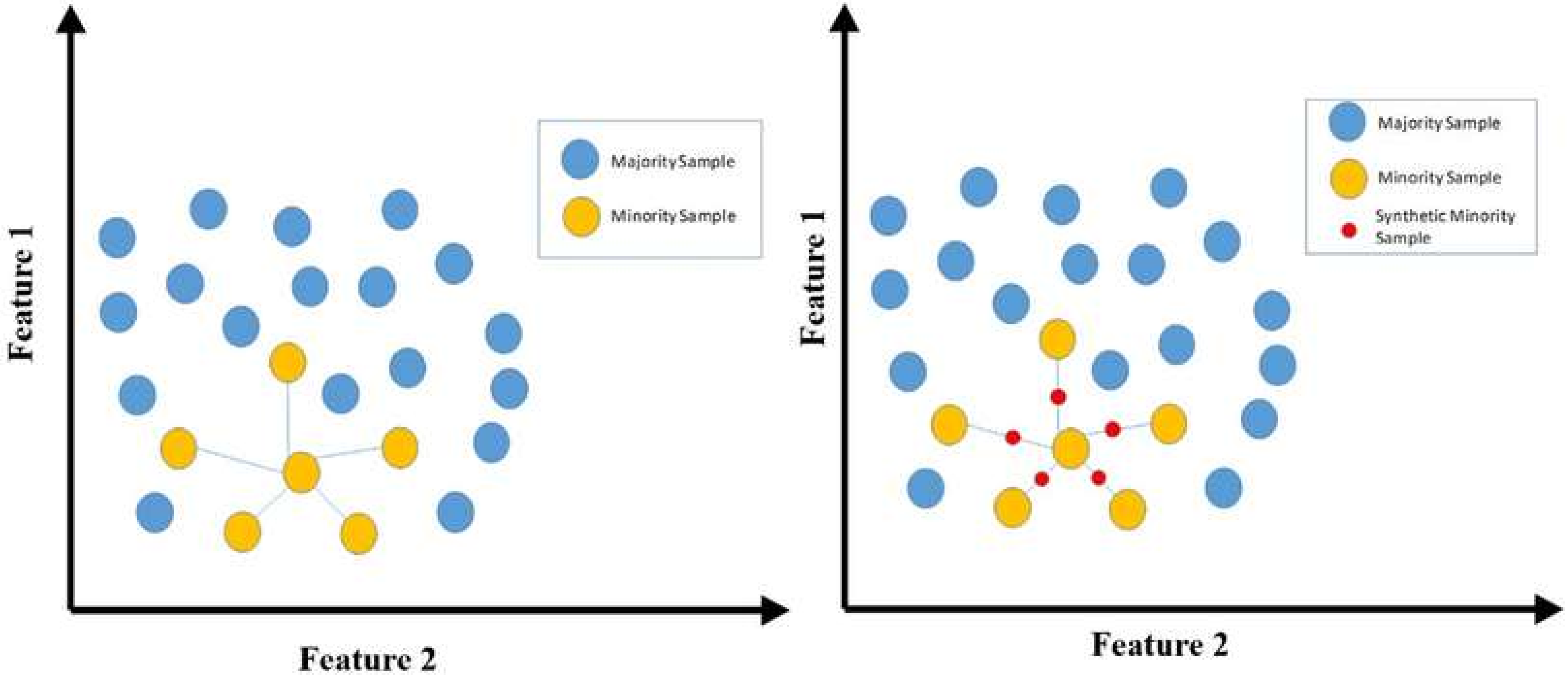
SMOTE: Synthetic Minority Over-sar

Nitesh V. Chawla
Department of Computer Science and Engineering, ENB 118
University of South Florida
4202 E. Fowler Ave.
Tampa, FL 33620-5399, USA

Kevin W. Bowyer
Department of Computer Science and Engineering
384 Fitzpatrick Hall
University of Notre Dame
Notre Dame, IN 46556, USA

Lawrence O. Hall
Department of Computer Science and Engineering, ENB 118
University of South Florida
4202 E. Fowler Ave.

HALL@CSEE.USF.EDU

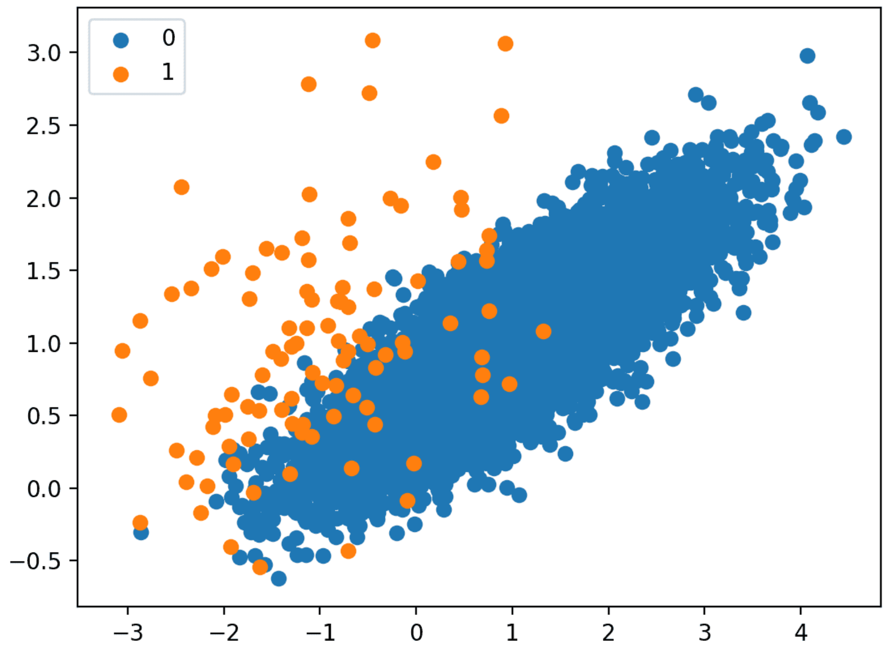


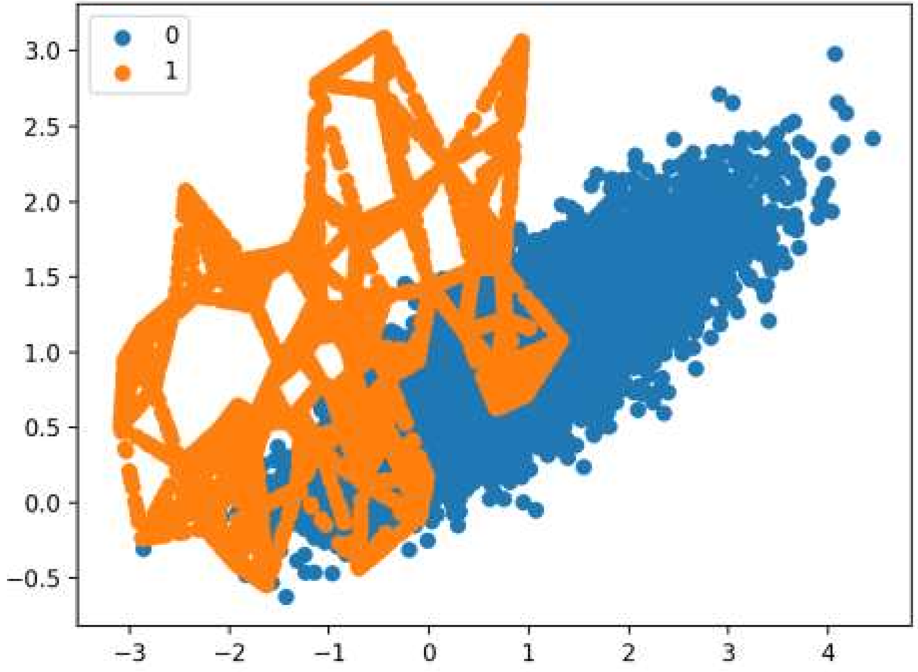
Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, *16*, 321-357.

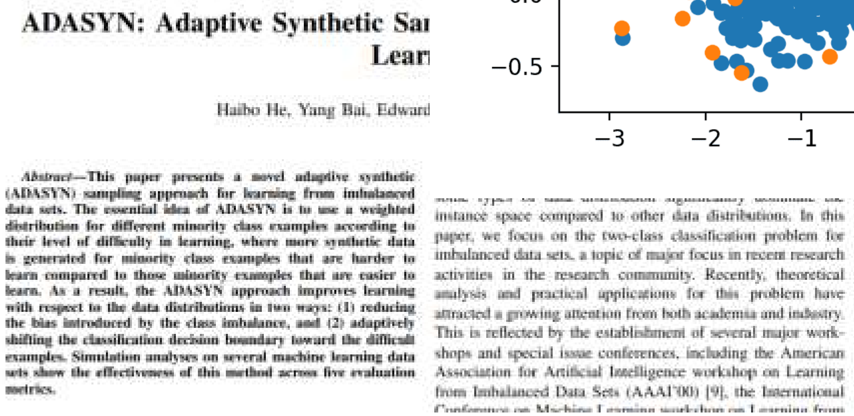
Handling imbalanced datasets

3. Synthetic oversampling of the dataset

Synthetic Minority Oversampling
Technique (SMOTE)







ADASYN: Adaptive Synthetic Sa
Lear

Haibo He, Yang Bai, Edward

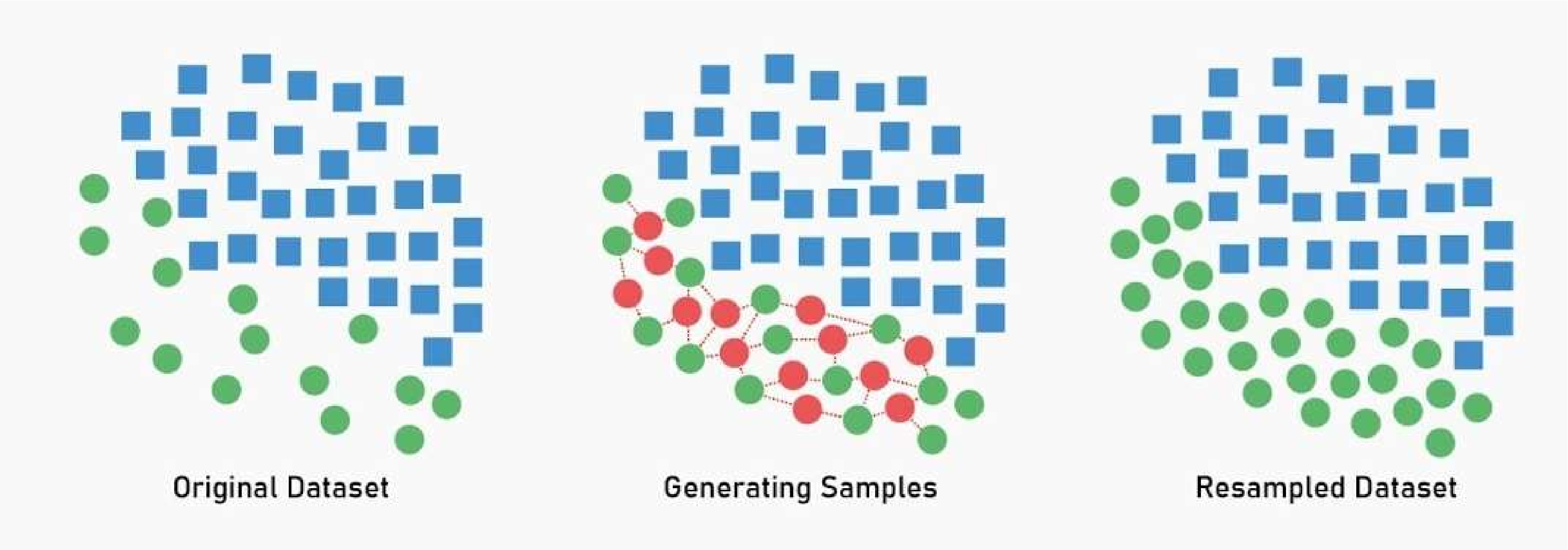
Abstract—This paper presents a novel adaptive synthetic (ADASYN) sampling approach for learning from imbalanced data sets. The essential idea of ADASYN is to use a weighted distribution for different minority class examples according to their level of difficulty in learning, where more synthetic data is generated for minority class examples that are harder to learn compared to those minority examples that are easier to learn. As a result, the ADASYN approach improves learning with respect to the data distributions in two ways: (1) reducing the bias introduced by the class imbalance, and (2) adaptively shifting the classification decision boundary toward the difficult examples. Simulation analyses on several machine learning data sets show the effectiveness of this method across five evaluation metrics.

instance space compared to other data distributions. In this paper, we focus on the two-class classification problem for imbalanced data sets, a topic of major focus in recent research activities in the research community. Recently, theoretical analysis and practical applications for this problem have attracted a growing attention from both academia and industry. This is reflected by the establishment of several major workshops and special issue conferences, including the American Association for Artificial Intelligence workshop on Learning from Imbalanced Data Sets (AAAI'00) [9], the International

He, H., Bai, Y., Garcia, E. A., & Li, S. (2008, June). ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)* (pp. 1322-1328). Ieee.

3. Synthetic oversampling of the dataset

Synthetic Minority Oversampling
Technique (SMOTE)



ADASYN: Adaptive Synthetic Sampling Appro
Learning

Haibo He, Yang Bai, Edwardo A. Garcia, and Shu

***Abstract*—This paper presents a novel adaptive synthetic (ADASYN) sampling approach for learning from imbalanced data sets. The essential idea of ADASYN is to use a weighted distribution for different minority class examples according to their level of difficulty in learning, where more synthetic data is generated for minority class examples that are harder to learn compared to those minority examples that are easier to learn. As a result, the ADASYN approach improves learning with respect to the data distributions in two ways: (1) reducing the bias introduced by the class imbalance, and (2) adaptively shifting the classification decision boundary toward the difficult examples. Simulation analyses on several machine learning data sets show the effectiveness of this method across five evaluation metrics.**

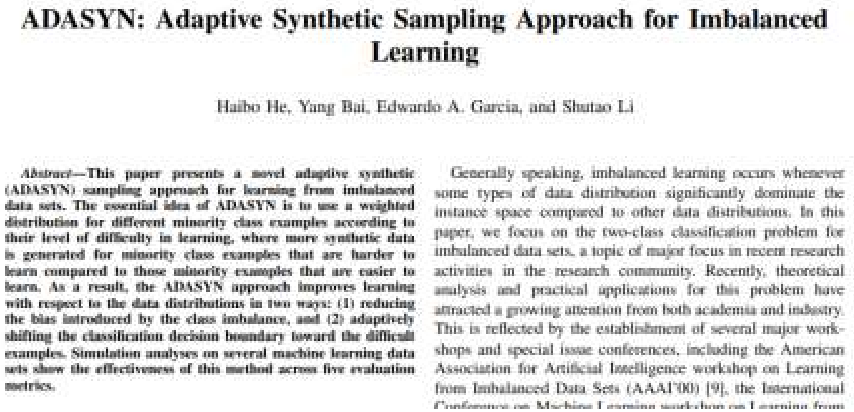
Generally speaking, imbalanced learning occurs whenever some types of data distribution significantly dominate the instance space compared to other data distributions. In this paper, we focus on the two-class classification problem for imbalanced data sets, a topic of major focus in recent research activities in the research community. Recently, theoretical analysis and practical applications for this problem have attracted a growing attention from both academia and industry. This is reflected by the establishment of several major workshops and special issue conferences, including the American Association for Artificial Intelligence workshop on Learning from Imbalanced Data Sets (AAAI'00) [9], the International

He, H., Bai, Y., Garcia, E. A., & Li, S. (2008, June). ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)* (pp. 1322-1328). Ieee.

3. Synthetic oversampling of the dataset

ADAptive SYNthetic oversampling
(Adasyn)

The same as SMOTE, but synthetic oversampling depends on the class distribution
-> it creates more synthetic samples near the boundary between the two classes (than within the distribution of the minority class)



ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning

Haibo He, Yang Bai, Edwardo A. Garcia, and Shutao Li

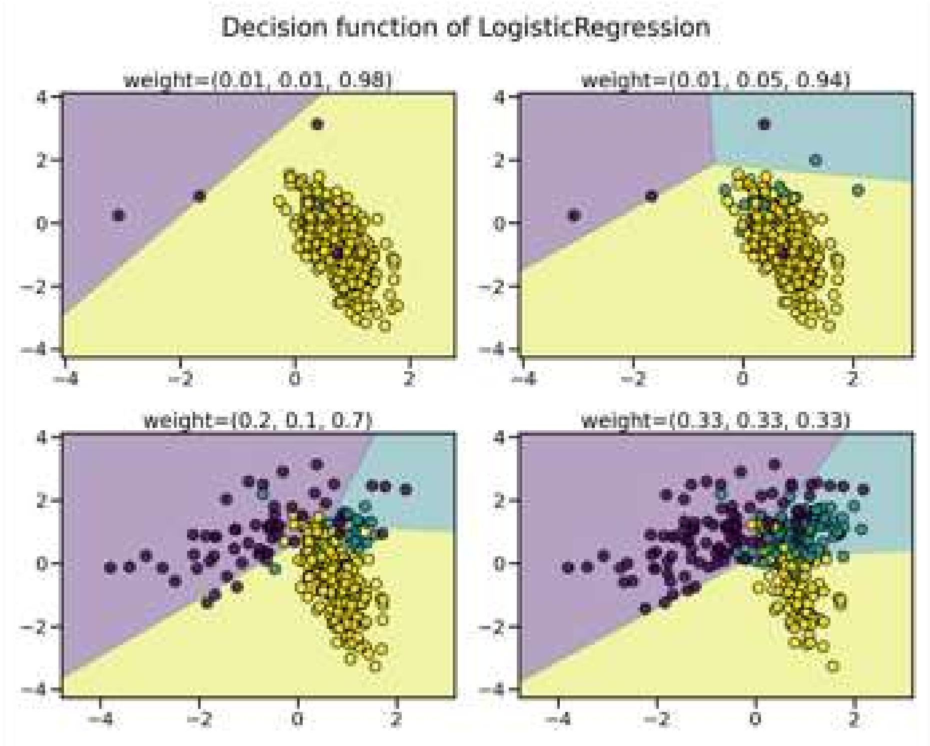
***Abstract*—This paper presents a novel adaptive synthetic (ADASYN) sampling approach for learning from imbalanced data sets. The essential idea of ADASYN is to use a weighted distribution for different minority class examples according to their level of difficulty in learning, where more synthetic data is generated for minority class examples that are harder to learn compared to those minority examples that are easier to learn. As a result, the ADASYN approach improves learning with respect to the data distributions in two ways: (1) reducing the bias introduced by the class imbalance, and (2) adaptively shifting the classification decision boundary toward the difficult examples. Simulation analyses on several machine learning data sets show the effectiveness of this method across five evaluation metrics.**

Generally speaking, imbalanced learning occurs whenever some types of data distribution significantly dominate the instance space compared to other data distributions. In this paper, we focus on the two-class classification problem for imbalanced data sets, a topic of major focus in recent research activities in the research community. Recently, theoretical analysis and practical applications for this problem have attracted a growing attention from both academia and industry. This is reflected by the establishment of several major workshops and special issue conferences, including the American Association for Artificial Intelligence workshop on Learning from Imbalanced Data Sets (AAAI'00) [9], the International

He, H., Bai, Y., Garcia, E. A., & Li, S. (2008, June). ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)* (pp. 1322-1328). Ieee.

3. Synthetic oversampling of the dataset

ADAptive SYNthetic oversampling
(Adasyn)



ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning

Haibo He, Yang Bai, Edwardo A. Garcia, and Shutao Li

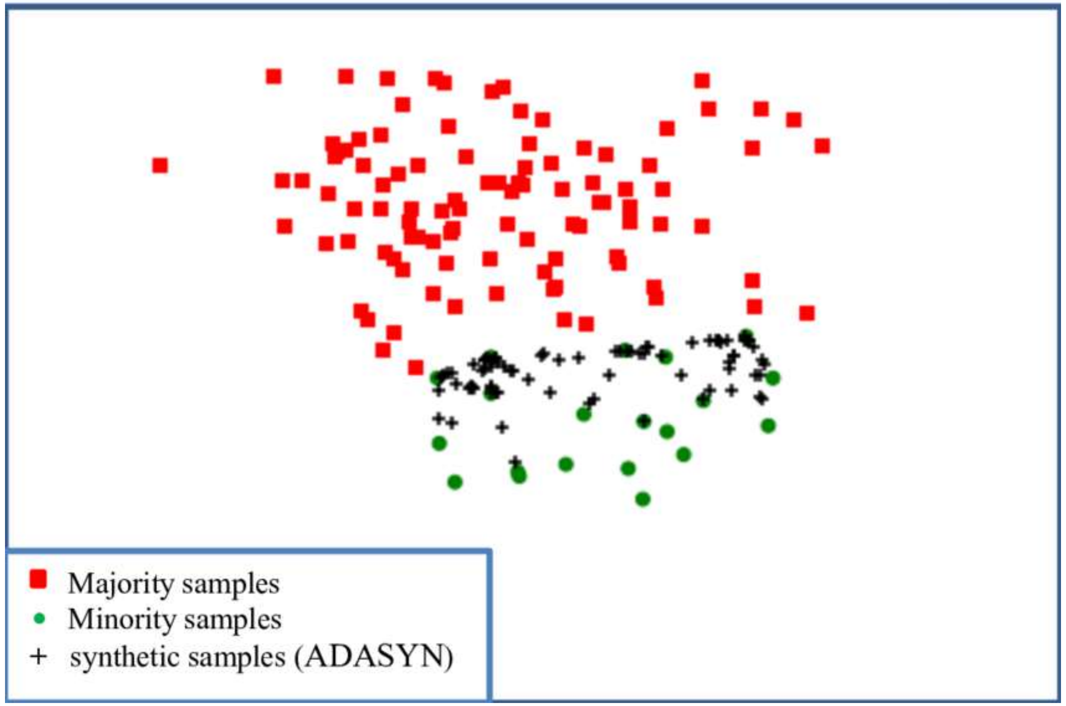
***Abstract*—This paper presents a novel adaptive synthetic (ADASYN) sampling approach for learning from imbalanced data sets. The essential idea of ADASYN is to use a weighted distribution for different minority class examples according to their level of difficulty in learning, where more synthetic data is generated for minority class examples that are harder to learn compared to those minority examples that are easier to learn. As a result, the ADASYN approach improves learning with respect to the data distributions in two ways: (1) reducing the bias introduced by the class imbalance, and (2) adaptively shifting the classification decision boundary toward the difficult examples. Simulation analyses on several machine learning data sets show the effectiveness of this method across five evaluation metrics.**

Generally speaking, imbalanced learning occurs whenever some types of data distribution significantly dominate the instance space compared to other data distributions. In this paper, we focus on the two-class classification problem for imbalanced data sets, a topic of major focus in recent research activities in the research community. Recently, theoretical analysis and practical applications for this problem have attracted a growing attention from both academia and industry. This is reflected by the establishment of several major workshops and special issue conferences, including the American Association for Artificial Intelligence workshop on Learning from Imbalanced Data Sets (AAAI'00) [9], the International Conference on Machine Learning workshop on Learning from

He, H., Bai, Y., Garcia, E. A., & Li, S. (2008, June). ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)* (pp. 1322-1328). Ieee.

3. Synthetic oversampling of the dataset

ADAptive SYNthetic oversampling
(Adasyn)



ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning

Haibo He, Yang Bai, Edwardo A. Garcia, and Shutao Li

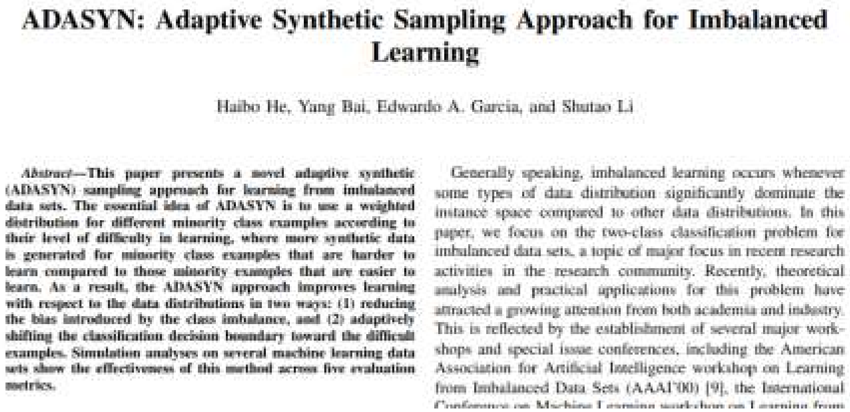
***Abstract*—This paper presents a novel adaptive synthetic (ADASYN) sampling approach for learning from imbalanced data sets. The essential idea of ADASYN is to use a weighted distribution for different minority class examples according to their level of difficulty in learning, where more synthetic data is generated for minority class examples that are harder to learn compared to those minority examples that are easier to learn. As a result, the ADASYN approach improves learning with respect to the data distributions in two ways: (1) reducing the bias introduced by the class imbalance, and (2) adaptively shifting the classification decision boundary toward the difficult examples. Simulation analyses on several machine learning data sets show the effectiveness of this method across five evaluation metrics.**

Generally speaking, imbalanced learning occurs whenever some types of data distribution significantly dominate the instance space compared to other data distributions. In this paper, we focus on the two-class classification problem for imbalanced data sets, a topic of major focus in recent research activities in the research community. Recently, theoretical analysis and practical applications for this problem have attracted a growing attention from both academia and industry. This is reflected by the establishment of several major workshops and special issue conferences, including the American Association for Artificial Intelligence workshop on Learning from Imbalanced Data Sets (AAAI'00) [9], the International

He, H., Bai, Y., Garcia, E. A., & Li, S. (2008, June). ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)* (pp. 1322-1328). Ieee.

3. Synthetic oversampling of the dataset

ADAptive SYNthetic oversampling
(Adasyn)



ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning

Haibo He, Yang Bai, Edwardo A. Garcia, and Shutao Li

Abstract—This paper presents a novel adaptive synthetic (ADASYN) sampling approach for learning from imbalanced data sets. The essential idea of ADASYN is to use a weighted distribution for different minority class examples according to their level of difficulty in learning, where more synthetic data is generated for minority class examples that are harder to learn compared to those minority examples that are easier to learn. As a result, the ADASYN approach improves learning with respect to the data distributions in two ways: (1) reducing the bias introduced by the class imbalance, and (2) adaptively shifting the classification decision boundary toward the difficult examples. Simulation analyses on several machine learning data sets show the effectiveness of this method across five evaluation metrics.

Generally speaking, imbalanced learning occurs whenever some types of data distribution significantly dominate the instance space compared to other data distributions. In this paper, we focus on the two-class classification problem for imbalanced data sets, a topic of major focus in recent research activities in the research community. Recently, theoretical analysis and practical applications for this problem have attracted a growing attention from both academia and industry. This is reflected by the establishment of several major workshops and special issue conferences, including the American Association for Artificial Intelligence workshop on Learning from Imbalanced Data Sets (AAAI'00) [9], the International

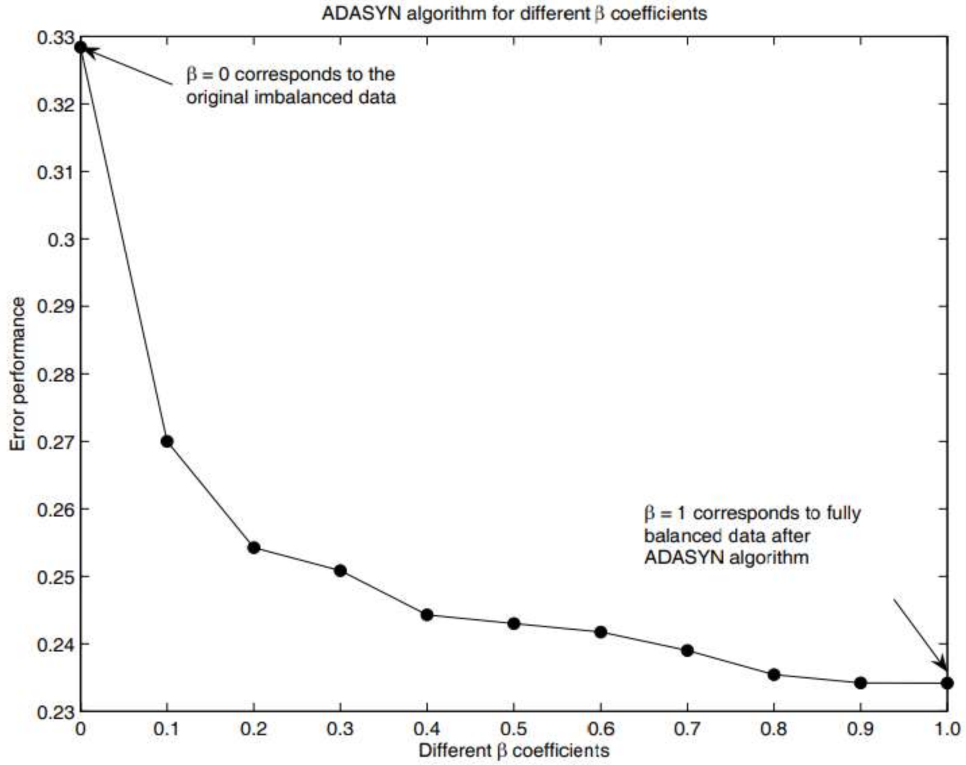


Fig. 1. ADASYN algorithm for imbalanced learning

He, H., Bai, Y., Garcia, E. A., & Li, S. (2008, June). ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)* (pp. 1322-1328). Ieee.

4. "Forcing" the behaviour of the ML model

i.e. cost-sensitive SVM

Specifically, each example in the training dataset has its own penalty term (*C* value) used in the calculation for the margin when fitting the SVM model. The value of an example's *C*-value can be calculated as a weighting of the global *C*-value, where the weight is defined proportional to the class distribution.

- C_i = weight_i * C

A larger weighting can be used for the minority class, allowing the margin to be softer, whereas a smaller weighting can be used for the majority class, forcing the margin to be harder and preventing misclassified examples.

- **Small Weight**: Smaller *C* value, larger penalty for misclassified examples.
- **Larger Weight**: Larger *C* value, smaller penalty for misclassified examples.

This has the effect of encouraging the margin to contain the majority class with less flexibility, but allow the minority class to be flexible with misclassification of majority class examples onto the minority class side if needed.

> *That is, the modified SVM algorithm would not tend to skew the separating hyperplane toward the minority class examples to reduce the total misclassifications, as the minority class examples are now assigned with a higher misclassification cost.*

— Page 89, Imbalanced Learning: Foundations, Algorithms, and Applications, 2013.