Working with imbalanced data in 2024

Train In Data

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.
Tampa, FL 33620-5399, USA

W. Philip Kegelmeyer WPK@CALIFORNIA.SANDIA.GOV
Sandia National Laboratories
Biosystems Research Department, P.O. Box 969, MS 9951
Livermore, CA, 94551-0969, USA

Home
Questions
Tags
Users
Jobs
Companies
Unanswered
TEAMS
overflow AI
Now available on Stack Overflow for Teams! AI features where you work: search, IDE, and chat.
Learn more
Explore Teams

Doesn't over(/under)sampling an imbalanced dataset cause issues?

Asked 3 years, 3 months ago Modified 3 years, 2 months ago Viewed 2k times

7

I'm reading a lot about how to use different metrics specifically for imbalanced datasets (e.g. two classes present, but 80% of the data is one class) and how to tackle the issue of imbalanced datasets.

One trick is to oversample, so to take more (or even duplic
underrepresented class. I've tried this and did achieve bett
easily just predict a single class for everything, achieving 8

However, I was wondering, will this model work well with r
science/machine learning is that your training data has to l
live data you're intending to use your model on. However,
that's 50% one class and 50% other, as opposed to the "na
class and 20% of the other.

So I guess the question in short is: Will oversampling my ir
distribution to 50/50 class distribution impact the usability

classification class-imbalance imbalanced-data

Data Science

Home
Questions
Tags
Users
Jobs
Companies
Unanswered
TEAMS
overflow AI
Now available on Stack Overflow for Teams! AI features where you work: search, IDE, and chat.
Learn more
Explore Teams

Why SMOTE is not used in prize-winning Kaggle solutions?

Asked 2 years, 7 months ago Modified 4 months ago Viewed 4k times

14

Synthetic Minority Over-sampling Technique SMOTE, is a well known method to tackle imbalanced datasets. There are many papers with a lot of citations out-there claiming that it is used to boost accuracy in unbalanced data scenarios.

But then, when I see Kaggle competitions, it is rarely used, to the best of my knowledge there are no prize-winning Kaggle/ML competitions where it is used to achieve the best solution. **Why SMOTE is not used in Kaggle?**

I even see applied research papers (where there are millions of $ at stake) that SMOTE is not used: Practical Lessons from Predicting Clicks on Ads at Facebook

Is this because it's not the best strategy possible? Is it a research niche with no optimal real-life application? Is there any ML competition with a high-reward where this was used to achieve the best solution?

I guess I am just hesitant that creating synthetic data actually helps.

machine-learning class-imbalance kaggle smote

abhishek ✓
@abhi1thakur

Question for machine learning twitter :)
Has oversampling techniques like SMOTE ever helped you in real-world scenarios?

2:02 PM · Jan 10, 2022

105 41 382 88

merve ✓ @mervenoyann · Jan 10, 2022
if you don't have a nice legit separation inside (which assigning a weight would work instead of SMOTE anyway) it doesn't work
I tried it two years ago in my old job, data was very imbalanced
I know that banks use it (at least banks here)

2 2 16

Konrad Banachewicz @UniBananStates · Aug 31, 2023
Never. Not a single time, and for a while I did keep on trying.

2 1 3 874

Bob Muenchen @BobMuenchen · Nov 13, 2018
Doing machine learning with unbalanced classes? SMOTE that data!
rdrr.io/cran/UBL/man/s... #rstats #machinelearning #statistics #PredictiveAnalytics

1 1

Frank Harrell
@f2harrell

Bob please don't do this. SMOTE is needed only because of the use of an improper discontinuous accuracy scoring rule. Don't mess with the data. Use the right stat methods.

8:20 PM · Nov 13, 2018

Carlos Mougan @CarlosMougan · Dec 28, 2021
Why SMOTE is not used in prize-winning Kaggle solutions? @kaggle

datascience.stackexchange.com
Why SMOTE is not used in prize-winning Kaggle solut
Synthetic Minority Over-sampling Technique SMOTE, is a well known method to tackle ...

1 8 26

Carlos Mougan @CarlosMougan · Dec 28, 2021
I would love to know the thoughts of some Kaggle Grandmasters, would be a nice xmas present. @kagglingdieter @tunguz @JFPuget

2 1 4

JFPuget 🇺🇦
@JFPuget

Because it does not work.

If you disagree then I welcome an example of SMOTE working. Not using synthetic data, real data only.

11:04 AM · Dec 28, 2021

To SMOTE, or not to SMOTE?

Yotam Elor
yotame@amazon.com
Amazon
New York, USA

Hadar Averbuch-Elor
hadarelor@cornell.edu
Cornell University
New York, USA

ABSTRACT

Balancing the data before training a classifier is a popular technique to address the challenges of imbalanced binary classification in tabular data. Balancing is commonly achieved by duplication of minority samples or by generation of synthetic minority samples. While it is well known that balancing affects each classifier differently, most prior empirical studies did not include strong state-of-the-art (SOTA) classifiers as baselines. In this work, we are interested in understanding whether balancing is beneficial, particularly in the context of SOTA classifiers. Thus, we conduct

predicted labels (*e.g.*, f_β, balanced-accuracy and Jaccard similarity coefficient).

Machine learning classifiers typically optimize a symmetric objective function which associates the same loss with minority and majority samples. Thus, considering imbalanced classification problems, as previously noted in [4, 18], there seems to be a discrepancy between the classifier's symmetric optimization and the non-symmetric metric of interest which might result in an undesirable bias in the final trained model.

On the other hand, theoretical investigations into proper metrics

SMOTE – tested with bad design

Most articles about SMOTE:

- Used weaker learners (decision trees, random forests, adaboost, SVMs and MLPs)
- Optimized metrics that require a threshold, like precision and recall, where the threshold was arbitrarily set at 0.5
  - Terrible design

Probability threshold

A probability threshold of 0.5 is only useful (at best) for balanced datasets.

With imbalanced data, we need to find the best threshold for the metric we want to optimize.

Apparently, the use of 0.5 is quite wide-spread.

To SMOTE, or not to SMOTE?

Yotam Elor
yotame@amazon.com
Amazon
New York, USA

Hadar Averbuch-Elor
hadarelor@cornell.edu
Cornell University
New York, USA

ABSTRACT

Balancing the data before training a classifier is a popular technique to address the challenges of imbalanced binary classification in tabular data. Balancing is commonly achieved by duplication of minority samples or by generation of synthetic minority samples. While it is well known that balancing affects each classifier differently, most prior empirical studies did not include strong state-of-the-art (SOTA) classifiers as baselines. In this work, we are interested in understanding whether balancing is beneficial, particularly in the context of SOTA classifiers. Thus, we conduct

predicted labels (*e.g.*, f_β, balanced-accuracy and Jaccard similarity coefficient).

Machine learning classifiers typically optimize a symmetric objective function which associates the same loss with minority and majority samples. Thus, considering imbalanced classification problems, as previously noted in [4, 18], there seems to be a discrepancy between the classifier's symmetric optimization and the non-symmetric metric of interest which might result in an undesirable bias in the final trained model.

On the other hand, theoretical investigations into proper metrics

To SMOTE or not to SMOTE

They tested SMOTE using:

- Several machine learning models, including xgboost, catboost and lightGBMs
- Probability based metrics like log loss and AUC and threshold metrics like precision and recall, with or without adjusting the threshold

To SMOTE or not to SMOTE

They found that:

- xgboost, catboost and lightGBMs outperform all other models
- SMOTE did not improve the performance of the GBMs
- SMOTE improved the performance of weak learners when evaluating probability based metrics
- SMOTE improved the performance of ALL classifiers when using threshold based metrics set at 0.5
- The improvement was lost when the threshold was properly adjusted.

Stop Oversampling for Class Imbalance Learning: A Critical Review

Ahmad B. Hassanat[a], Ahmad S. Tarawneh[b], Ghada A. Altarawneh[c], Abdullah Almuhaimeed[d]

[a] *Faculty of Information Technology, Mutah University, Karak, Jordan*
[b] *Dept. of Algorithms and Their Applications, Eötvös Loránd University, Budapest, Hungary*
[c] *Dept. Accounting, Mutah University, Karak, Jordan*
[d] *The National Centre for Genomics and Bioinformatics, King Abdulaziz City for Science and Technology, Riyadh 11442, Saudi Arabia*

Synthetic data created by SMOTE are not representative of the real data.

In other words, SMOTE creates observations that do not exist in real life.

THE HARMS OF CLASS IMBALANCE CORRECTIONS FOR MACHINE LEARNING BASED PREDICTION MODELS: A SIMULATION STUDY.

A PREPRINT

Alex Carriero
Julius Center for Health Sciences and Primary Care
University Medical Center Utrecht
Netherlands
a.j.carriero@umcutrecht.nl

Kim Luijken
Julius Center for Health Sciences and Primary Care
University Medical Center Utrecht
Netherlands

Resampling methods may affect classifier calibration, irreversibly.

To SMOTE or not to SMOTE?

- If you are using strong GBMS, most likely SMOTE will not improve model performance.
- SMOTE can improve the performance of some weak learners
- SMOTE can improve the performance in cases where the classifiers do not output probabilities

How to approach imbalanced data

- Use strong GBMs whenever possible (xgboost and Catboost)
- Adjust the probability threshold when evaluating models with threshold dependent metrics like precision and recall
- Use more than 1 metric (resampling has been shown to improve the values of some metrics but deteriorate others)
- Explore the result of your data manipulation (does the synthetic data make sense)
- Check the calibration of your classifiers.

Train In Data

THANK YOU

www.trainindata.com