<https://jair.org/index.php/jair/article/view/11192/26406>

[SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary](https://jair.org/index.php/jair/article/view/11192)

* In the 1990s as more data and applications of machine learning and data mining started to become prevalent, an important challenge emerged: how to achieve desired classification accuracy when dealing with data that had significantly skewed class distributions
* n many cases, the specificityor local accuracy on the majority class examples overwhelmed the one achieved on the minority ones.
* Undersampling
  + t may also might discard someuseful examples for the modeling of the classifier. Particularly when the ratio of imbalanceis high, then more examples need to be removed leading to the problem of lack of data(Wasikowski & Chen, 2010). This may affect the generalization ability of the classifier.
* In 2002, Chawla, Bowyer, Hall, and Kegelmeyer (2002) proposed a novel approach as analternative to the standard random oversampling. The idea was to overcome the overfittingrendered by simply oversampling by replication, and assist the classifier to improve itsgeneralization on the testing data. Instead of “weighting” data points, the basis of this newdata preprocessing technique was to create new minority instances. This technique wastitled Synthetic Minority Oversampling Technique, now widely known as SMOTE (Chawlaet al., 2002). The basis of the SMOTE procedure was to carry out an interpolation amongneighboring minority class instances. As such, it is able to increase the number of minorityclass instances by introducing new minority class examples in the neighborhood, therebyassisting the classifiers to improve its generalization capacity
  + Oversampling by replication not done because already have danger of overfitting small low income dataset
* However, that celebration was short-lived. Hequickly realized that merely by guessing majority class, he would have achieved an accuracyof 97.68% (which was the majority class distribution in the original data).
  + Particularly harmful when predicting MMR, as could hide stark medical reality
* First, the total amount of oversamplingN(aninteger value) is set up, which can either be set-up to obtain an approximate 1:1 classdistribution or discovered via a wrapper process (Chawla et al., 2008). Then, an iterativeprocess is carried out, composed of several steps. First, a minority class instance is selectedat random from the training set. Next, itsKnearest neighbors (5 by default) are obtained.Finally,Nof theseKinstances are randomly chosen to compute the new instances by interpolation.
* The first extensions of SMOTE motivatedby its well known drawback of generating overlapped and noisy examples was the ad-dition of a noise filtering step just after SMOTE process ends
* The imbalancelearning correspondence for regression tasks is the correct prediction of rare extreme val-ues of a continuous target variable. In the work of Torgo et al. (2015), several techniquesfor resampling were successfully applied for regressio
* n bioinformatics problems, it is usual to have high-dimensional classification prob-lems. In the work of Blagus and Lusa (2013), SMOTE was tested in such scenariosin both theoretical and empirical perspectives. most important was that SMOTE has hardly any effect on most classifiers trained onhigh-dimensional data. Other techniques such as Undersampling may be preferableon high-simensional settings.
* When working in the scenario of imbalanced classification, we must be aware that the skewedclass distribution is not the only drawback for the performance degradation
* We refer to a dataset containing small disjuncts when some concepts (disregard their class)are represented within small clusters
  + Inthe case of imbalanced classes, this problem occurs very often as underrepresented conceptsare usually located in small areas of the dataset
* The problem of small disjuncts affects to a higher degree, those learning algorithmswhose procedure is based on a divide-and-conquer strategy. Since the original problemis divided into different subsets, in several iterations this can lead to data fragmentation(Friedman, 1996). Some clear examples of this behavior are decision trees
  + Prevents generalisation because maybe have a handful of samples in leaf
  + his synergy is straightforward as information is barely represented inthose small disjuncts. Therefore, learning classifiers cannot carry out a good generalizationwhen there is not enough data to represent the boundaries of the problem (Jo & Japkowicz,2004; Wasikowski & Chen, 2010). This way, small disjuncts, noisy data and lack of dataare three inter-related problems that comprise a challenge to the research community inimbalanced classification
  + Another approach is to apply a synergy of preprocessing models, i.e. filtering and/orinstance generation to remove those instances that are actually noisy prior to the SMOTEapplication (S ́aez et al., 2015; Verbiest, Ramentol, Cornelis, & Herrera, 2014). Some studiesshown that simple undersampling techniques such as random undersampling and cleaningtechniques are known to be robust for different levels of noise and imbalance (Seiffert et al.,2014). This way, many hybrid approaches between filtering techniques and SMOTE havebeen developed so far, since this allow to improve the quality of the data either a priori (fromthe original data), a posteriori (from the preprocessed data) or iteratively while creatingnew synthetic instances
* Among all data intrinsic characteristics, the overlapping between classes is possibly themost harmful issue (Garc ́ıa et al., 2008). It is defined as those regions of the data spacein which the representation of the classes is similar. This situation leads to develop aninference with almost the same a priori probabilities in this overlapping area, which makesvery hard or even impossible the distinction between the two classes
* In addition to the former, we must take into account that the dimensionality problemalso gives rise to the phenomenon of hubness (Radovanovic, Nanopoulos, & Ivanovic, 2010),defined as a small number of points that become most of the observed nearest neighbors
  + n the case of the SMOTE procedure, this affects the quality of the new synthetic examplesfor two inter-related reasons (Blagus & Lusa, 2013). On the one hand, the computation ofthe neighborhood becomes skewed to the actual one. On the other hand, the variance forthe new created instances becomes higher
  + Another simpler solution is to ben-efit from the use of a feature selection approach prior to the application of the SMOTEoversampling, as suggested in several work
  + Finally, feature extraction to transform the problem into a lower dimensional space isanother way to address this issue. When this process is carried out before the applicationof SMOTE, the new clusters of this transformed dataset may allow a better generation ofinstances