ACTIVE SMOTE for Imbalanced Medical Data Classification

Raul Sena, Sana Ben Hamida

ACTIVE SMOTE for Imbalanced Medical Data Classification

Raul SENA[1] and Sana BEN HAMIDA[2]

[1] raul.sena-rojas@dauphine.com
[2] sana.mrabet@dauphine.psl.eu
Paris Dauphine University, PSL Research University
CNRS, UMR[7243], LAMSADE, France.

Abstract. Classifying imbalanced data is a big challenge for machine learning techniques, especially for medical data. To deal with this challenge, many solutions have been proposed. The most famous methods are based on the Synthetic Minority Over-sampling Technique (SMOTE), which creates new synthetic instances in the minority class. In this paper, we study the efficiency of the SMOTE-based methods on some imbalanced data sets. We then propose extending these techniques with Active Learning to control the evolution of the minority class better. Active Learning uses uncertainty and diversity sampling to choose wisely the data points from which the synthetic samples will be generated. To evaluate our approach, we make comprehensive experimental studies on two medical data sets for diabetes diagnosis and breast cancer diagnosis.

Keywords: Imbalanced medical data · Machine Learning · SMOTE · Active Learning · Diversity Sampling · Uncertainty Sampling · Diabetes Diagnosis · Breast Cancer Detection

1 Introduction

The main objective of machine learning applications in the medical field is to propose efficient diagnostic tools with null or very low error probability. Nowadays, the availability of data is no more a difficulty. The amount of information available allows to train different types of learners. However, medical data is often extremely imbalance, where, minority class (known as the "positive" class) are far less than majority classes (known as the "negative" class).

The main difficulty with imbalanced data is that the classification algorithm could be biased towards the majority classes. This Bias induces a higher misclassification rate in the minority class [3]. This problem takes on other dimensions with medical data because classification is applied to generate models for diagnosing some diseases such as cancer, diabetes, etc. In this case, bias in the diagnostic models is not tolerated. Indeed, in medical diagnostics, mislabeling a patient as a healthy individual is expensive and often can lead to deadly consequences. Thus, addressing the class imbalance for medical data is crucial for machine learning tasks [20,17,5].

Many solutions have been proposed to deal with the class imbalance problem in Machine Learning, that could be classified into three categories: Cost-sensitive methods, algorithmic modification methods, and data pre-processing. The last method uses either the under or over sampling techniques, that eliminate or replicate instances until the classes are balanced. Data pre-processing includes also the Synthetic Minority Over-sampling Technique (SMOTE) and its variation, which achieves the same purpose by creating new synthetic instances from the minority class [7]. The efficiency of each approach depends on the context. For medical diagnostics, using the under/oversampling could induce a loss of information in the training sample. Moreover, with a high imbalance ratio, the synthetic samples generated by SMOTE in the positive class could overcome the original samples relative to patients diagnostic as positive cases.

In this paper, we propose a novel scheme to solve the imbalanced data problem for medical diagnostic. This scheme combines SMOTE with Active Learning, and we call it Active SMOTE. It proposes a twofold contribution. First, instead of choosing a sample at random from the training set to use as the pivot point to generate the synthetic samples (as classical SMOTE does), Active SMOTE chooses the points intelligently with uncertainty and diversity sampling, which are two techniques of Active Learning. The training sample is balanced progressively in incremental steps, that we call training epochs. Thus, at each step, the current synthetic samples could be used with the original samples to generate new synthetic samples in the minority class.

This paper is organized as follows. In section 2, we summarize the previous solutions to solve the class imbalance problem in Machine Learning (ML). Section 3 introduces the Active SMOTE method and details how SMOTE is combined with Active Learning. Section 4 explains the methodology used to compare our proposed algorithm with other sampling techniques. In section 5, we evaluate with some graphs and figures the performance of our proposed algorithm. Finally, section 6 makes general conclusions and gives some research ideas to continue forward.

2 Handling Class Imbalance for Classification - Some Related Works

2.1 Cost Sensitive Learning

Cost-sensitive learning is an aspect of algorithm-level modifications for class imbalance. This modification refers to a specific set of algorithms sensitive to different costs associated with certain characteristics of considered problems. These costs can be learned during the classifier training phase or be provided by a domain expert. There exist two different views on cost-sensitive learning in the literature. These are the following:

1. Cost associated with classes: This technique considers that making errors on instances coming from a particular class is associated with a higher cost [18]. There are two views for this approach: A financial perspective (e.g.,

giving credit to a person with a bad credit score will potentially cause higher losses to a bank than declining credit to a person with a good score) or the other scenario priority/health/ethical issues (e.g., sending a cancer patient home is much more costly than assigning additional tests to a healthy person). In general, the misclassification cost of the minority examples must be higher than that of the majority examples [2].

2. Cost associated with features: This method supposes that obtaining a particular feature is connected to a given cost, also known as test cost. We can view this from a monetary perspective (e.g., a feature is more expensive to obtain as it requires more resources) or other inconveniences (e.g., the measurement procedure is unpleasant, puts a person at risk, or is difficult to obtain). In other words, this approach aims at creating a classifier that obtains the best possible predictive performance while utilizing features that can be obtained at the lowest cost possible.

2.2 Data Level Preprocessing Methods

Data preprocessing methods consist of procedures to modify the imbalanced dataset to a more adequate or balanced data distribution [9]. This is helpful for many classifiers because rebalancing the dataset significantly improves their performance. This subsection will review the undersampling and oversampling techniques such as SMOTE. These techniques are simple and easy to implement. However, no clear rule tells us which technique works best. Resampling techniques can be categorized into three groups:

1. Oversampling methods: This method replicates some instances or creates new instances from existing ones, thus creating a superset of the original dataset.

2. Undersampling methods: Create a subset of the original dataset by eliminating instances (usually negative class instances).

3. Hybrid methods: A combination of Oversampling and Undersampling techniques.

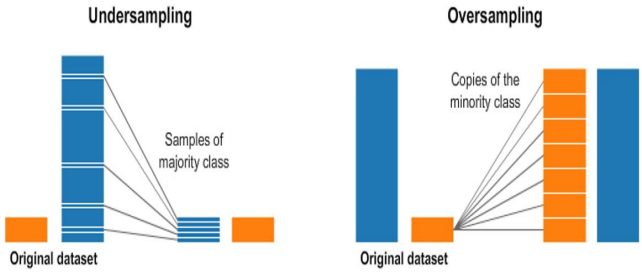


Fig. 1. Random Under and Over Sampling

Random Under and Over Sampling There are many ways to implement the previous techniques, where the simplest preprocessing are non-heuristic meth-

ods like random undersampling and random oversampling, as shown in Figure 1. Nevertheless, these techniques have some drawbacks. In the case of undersampling, the major problem is that it can discard potentially valuable data that could be used in the training process, reducing our dataset's variability. On the other hand, our classifier can occur overfitting with random oversampling because it makes exact copies of existing instances.

To tackle the previous problems, more sophisticated methods have been proposed. The "Synthetic Minority Oversampling Technique" (SMOTE) has gained popularity among them. In short, its main idea is to overcome overfitting posed by random oversampling with the generation of new instances with the help of interpolating between the positive instances closer to each other. However, SMOTE could generate noise samples, boundary samples and overlapping samples [19]. Thus, many variants have been proposed, such as Borderline SMOTE and ADASYN, that we present below since they are used in the experimental study.

SMOTE: Synthetic Minority Oversampling Technique As stated before, the problem of random oversampling is that because it replicates the exact copies of existing instances, no new information is added to the model's training; therefore, there is a high risk of overfitting. Here is where SMOTE comes in handy because instead of applying a simple replication of the minority class, the central idea of SMOTE is to generate new synthetic samples. This procedure focuses on the "feature space" rather than the "data space" since these new examples are created by interpolating several minority class instances closer. The process to generate new instances with SMOTE is shown in Figure 2.

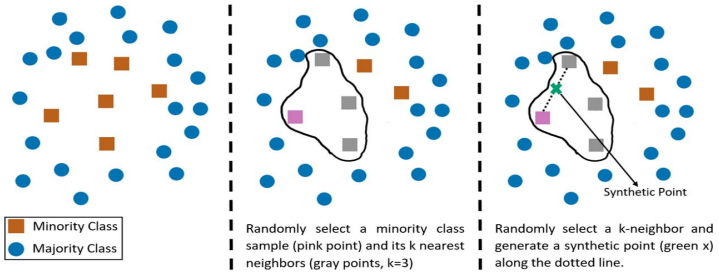


Fig. 2. SMOTE algorithm

Borderline SMOTE Many Machine Learning algorithms like Logistic Regression and Support Vector Machines use the concept of decision boundary to decide whether an example belongs to one class or another. This decision boundary tries to learn the limits of each class as accurately as possible in the training process. Then when the decision boundary is set, if an example lies far away from it, there is a small probability that it will belong to the opposite class. It is as if the decision boundary divides the space into regions where each region belongs to one class.

Based on the previous statement, the algorithms state that examples away from the borders may contribute little to classification. With this in mind, a new method of oversampling minority class examples was proposed, Borderline-SMOTE [12], in which only the limited examples of the minority class will be over sampled (the ones close to the decision boundary). It is important to clarify that the points close to the borderline are more important, but there is greater uncertainty about which class they belong to, so it is riskier to create synthetic points there. This method differs from the existing ones of oversampling, in which all minority examples or a random subset of the minority class are oversampled [15].

Borderline SMOTE knows which points are closer to the borderline because it classifies them into three categories, as shown in Fig 3. Then it only uses the danger points to generate synthetic samples.

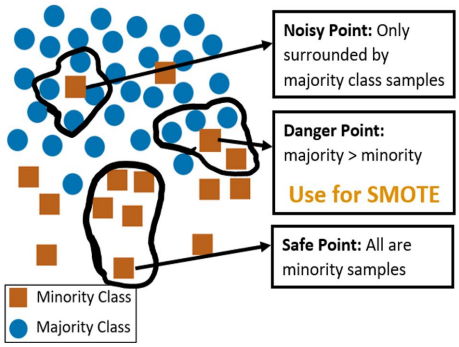


Fig. 3. Borderline Smote

ADASYN Adaptive synthetic sampling (ADASYN) is an attempt to enhance the SMOTE performance by modifying minority instance selection. It adaptively changes the number of artificial minority examples according to the density of the majority instances around the original minority ones [13]. It reduces the bias introduced by the class imbalance and adaptively shift the classification decision boundary toward the difficult instances.

Other SMOTE based methods SMOTE is the method that has received the most interest from researchers in addressing the imbalance problem. Several techniques were then proposed based on SMOTE with very close objectives. The main purpose of these variants is to avoid noise in the generated training sample. These methods use either techniques to study the disparity and density of the data, or clustering methods to identify safe samples. They can also be extended with Ensemble learning approaches. For example, in RSMOTE (Robust SMOTE)[4], relative density has been introduced to measure the local density of every minority sample, and the non-noisy minority samples are divided into the borderline samples and safe samples adaptively basing their distinguishing characteristics of relative density.

Ma and Fan [20] proposed CURE-SMOTE (Clustering Using Representatives-based Synthetic Minority Over-sampling TEchnique) that uses clustering to sample the training data. Then, SMOTE is performed on the revealed samples. Similarly, Xu et al.[16] proposed KNSMOTE combining k-means clustering with SMOTE in imbalanced medical data. KNSMOTE uses a k-means to cluster the instances and find so-called "safe samples" and remove noise. Then KNSMOTE creates synthetic samples based on founded "safe samples". Many other variants could be found in the literature such as DBSMOTE that uses density-based approach, MWMOTE that analyzes the most difficult minority examples, SMOTECSELM (Synthetic Minority Over-Sampling TEchnique based on Class-Specific Extreme Learning Machine) and many others. An extended review with a comprehensive analysis could be found in [10].

2.3 Algorithm Level Processing with Active Learning

Active Learning is another aspect of algorithm-level modifications. Active learning methods are used to select the instances to be considered in order to control the learning cost [1,11], mainly in the context of massive data, or to select the most informative instances in order to improve the quality of the obtained classifier. The active selection system is integrated in the iterative engine of the learner.

In the context of imbalanced data, active learning can be used to balance the training sample by selecting the most representative instances from the majority class [8], eliminating noisy examples from the minority class [21], and reducing the overall imbalance ratio. Active learning has been tested with iterative learning algorithms, essentially SVCs, particularly with not fully labeled data sets, and Genetic Programming as a scaling solution for classifying large imbalanced data [14].

Although active learning does not directly change the learning procedure, it is considered an algorithm-level solution. It is built into the learning process, unlike preprocessing approaches that are executed before learning begins [11].

The main purpose of Active Learning is to apply a dynamic data sampling to evolve the training data along the training process. The main question is how do we choose samples for a training set? What sample will increase the algorithm performance? At first, this problem may sound disconnected from the imbalanced class problem. However, in our case, the question is: What are the points from the minority class that we need to choose first to generate the synthetic samples, so we can finally have good model performance?

For this purpose, we have selected two sampling approaches: uncertainty sampling and diversity sampling.

2.4 Uncertainty Sampling:

The general idea of active learning is to iteratively provide the algorithm with new data, allowing it to improve the performance of the generated classification models. Intuitively, the instance selection method must be driven by the quality

of the obtained classifiers, that can be measured with the uncertainty. The more uncertain a prediction is, the more useful the selected instances will be for the learner. Uncertainty sampling is the set of techniques for identifying the least confident samples with the highest uncertainty near a decision boundary, to be inserted in the new training sample. It uses uncertainty measures for a classified item. There are many ways of measuring uncertainty, like least, margin or ratio of confidence, and entropy.

- Margin of confidence sampling: It computes the difference between the two most confident predictions.
- Least confidence sampling: It is defined by the difference between the most confident prediction and the maximum confidence (100%).
- Ratio of confidence sampling: Ratio between the two most confident predictions.
- Entropy-based sampling: Difference between all predictions, as defined by information theory. In our example, entropy-based sampling would capture how much every confidence differed from every other. In the binary classification problem, entropy base sampling is the same as the margin of confidence.

2.5 Diversity Sampling:

This type of sampling tackles the problem of identifying where the model might be confident but wrong due to undersampled or non-representative data. It is based on various data sampling approaches helpful in identifying gaps in the model's knowledge, such as clustering, representative sampling, and methods that identify and reduce real-world bias in the models. Collectively, these techniques are known as diversity sampling.

In the figures 4 and 5, we can see how uncertainty sampling chooses items closer to the borderline for different ML algorithms like Decision Trees, SVC, Random Forest, and Naive Bayes. In contrast, diversity sampling selects samples that differ from one another or, in other words, are more different.

3 Active SMOTE : Combining SMOTE with Active Learning - Proposed Algorithm

Active SMOTE aims to combine SMOTE with Active Learning. In other words, instead of choosing a point at random from the training set to use as the pivot point to generate the synthetic samples, we will choose the points intelligently with uncertainty and diversity sampling. Practically, we can say there are two main phases of the new proposed algorithm. An uncertainty sampling phase and a diversity sampling phase. These phases are shown in the following images:

The uncertainty sampling phase shown in Figure 4 serves to select the items that are close to the decision boundary. First, we train a machine learning model with all the data, then calculate the probability of belonging to the minority class of all minority class samples. Finally, we compute the uncertainty of the model

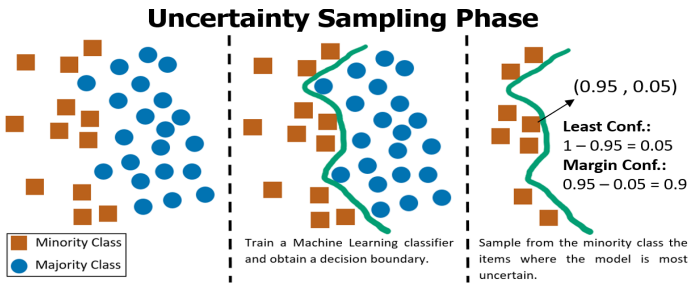


Fig. 4. Uncertainty Sampling Phase of SMOTE with Active Learning

base in an uncertainty measure and select a percentage of the most uncertain minority class samples. The percentage of the items that we are going to select is a hyperparameter. After that, we proceed to the diversity sampling phase shown in Figure 5.

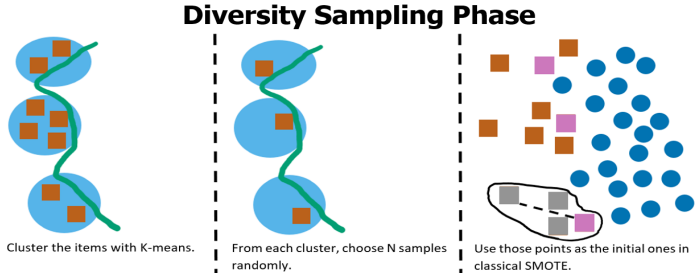


Fig. 5. Diversity Sampling Phase of SMOTE with Active Learning

The diversity sampling phase aims to make a diverse sample of the selected most uncertain items. First, we divide the most uncertain items in k clusters using K-means, and then we make a stratified sample from every cluster. The number of k clusters is a hyperparameter that needs to be tuned, as well as the number of items we will select from each cluster. We selected K-means as the clustering method because it is the most used clustering strategy. However, a good line of research could be to compare other clustering methods.

The different steps of Active SMOTE method are illustrated in algorithm 1. First, at each epoch, Active SMOTE apply the uncertainty sampling on the minority class sample $Xmin$ (lines 2 to 6). The confidence measure, margin or least, is a required parameter of the algorithm ($uncertain\_measure$). After retaining the appropriate portion ($perc_{uncertain}$) from the resulting sample $X_{uncertain}$ (line 8), this last one is clustered with K-means, and some points are selected from each cluster to obtain the final sample $X_{diversity}$ (lines 9 and 10). The third step consists on generating one synthetic sample from every diversity cluster in $X_{diversity}$ with the original SMOTE using $k = 5$ as number of neighbors to generate a new sample, and save it on the X_{smote} set (line 11). Finally,

the ML algorithm is trained on the over sampled set gathering X_{maj}, X_{min} and X_{smote}.

Algorithm 1 Active SMOTE Algorithm

Parameters:
X_{min}: Minority class sample
X_{maj}: Majority class sample
epochs: Number of iterations
$perc_{uncertain}$: % of most uncertain samples from T to retain (uncertainty sampling)
$perc_{diversity}$: % of samples from uncertainty sampling that are going to pass to diversity sampling
uncertain_measure : Uncertainty measure to use for confidence (margin or least confidence)
k: Number of clusters for k-means in diversity sampling
N: Number of samples points

for $i = 1$ to *epochs* **do**
  if *uncertain_measure*="margin confidence" **then**
    $X_{uncertain} \leftarrow$ compute model's margin confidence with X_{min}
  else
    $X_{uncertain} \leftarrow$ compute model's least confidence with X_{min}
  end if
  $X_{uncertain} \leftarrow$ sort $X_{uncertain}$ by descending order
  $X_{uncertain} \leftarrow$ retain $perc_{uncertain}$ samples from $X_{uncertain}$
  #Second do Diversity Sampling
  $X_{diversity}$ = Cluster $X_{uncertain}$ with K-means obtaining k clusters.
  $X_{diversity}$ = Do (cluster sampling of $X_{diversity}$) until size ($X_{diversity} * perc_{diversity}$) is reach
  $X_{smote} = SMOTE(X_{diversity}, N = 100, k = 5)$
  $X_{train} \leftarrow (X_{min} + X_{smote} + X_{maj})$
  model = retrain the model with the new training set X_{train}
end for
Return model

ML technique	Parameters
Logistic Regression	$0.001 \leq C \leq 2$
SVC	kernel : 'rbf' gamma (adjustment degree) $\in \{0.001, 0.01, 0.1, 1\}$ C (regulation) $\in \{1, 10, 50, 100, 200, 300, 1000\}$
Gradient Boosting	*n_estimator* $\in \{10, 20, 50\}$ learning rate $\in \{0.075, 0.01, 0.005\}$ max depth $\in \{1, 2, 3, 4, 5\}$

Table 1. ML methods' parameter setting

4 Experimentation

4.1 ML models evaluated

To evaluate the performance of Active SMOTE, we trained the following three ML classifiers: Logistic Regression (LR), Support Vector Machines (SVCs) and

Gradient Boost (GB). The training procedure involves performing a random search with ten-fold cross-validation using the hyperparameter space outlined in the table 1.

4.2 Sampling Techniques Evaluated

The training data was generated from the following two methods:

Classical sampling Methods The classical methods are SMOTE variants, simple random undersampling, and oversampling. All these techniques were used to make a balanced dataset, except with UNBAL. Otherwise, all oversampling techniques generate synthetic samples in the minority class with the specified strategy. A Python implementation of these techniques is available in the library "imbalanced-learn"[3] The abbreviations are the following:

- UNDER: Random undersampling of the majority class.
- OVER: Random oversampling of the minority class.
- SMOTE: Create synthetic samples with the original SMOTE.
- SVM: Generate synthetic samples with SVM SMOTE.
- BORDER: Create synthetic samples with Borderline SMOTE.
- ADASYN: Generate synthetic samples with Adasyn SMOTE.
- UNBAL: Do not change the original training dataset. That is, keeping the classes imbalanced.

New Proposed Methods The newly proposed methods are a combination of different Active Learning techniques with SMOTE, random (simple SMOTE) is used to evaluate if randomly choosing the pivot points has the same or better results as Active Learning with SMOTE. A more precise description of the newly proposed methods is the following:

- Random (simple SMOTE): At each epoch, randomly generate N synthetic samples with the original SMOTE.
- Diversity & Uncertainty – margin: At each iteration, generate N synthetic samples with the technique described in Algo. 1. In uncertainty sampling, margin confidence is used as an uncertainty measure.
- Diversity & Uncertainty – least: At each epoch, generate N synthetic samples with the technique described in Algo. 1. In uncertainty sampling, the least confidence is used as an uncertainty measure.
- Uncertainty – margin: At each epoch, generate N synthetic samples with the technique described in Algo. 1, only doing uncertainty sampling with margin confidence.

[3] `https://imbalanced-learn.org/stable/index.html`

4.3 Data Sets Evaluated

Two medical data sets (obtained from the public UCI data repository[4]) are used to assess the performance of Active SMOTE. The first one is the PIMA data set, which the target is to predict whether or not a patient will have diabetes based on specific diagnostic measurements included in the dataset. It consists of 768 females aged 21 or above, with 500 negative and 268 positive instances.

The second data set is "Breast Cancer Wisconsin" used in [6]. Features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe the characteristics of the cell nuclei present in the image. The target is to identify if a tumor is benign or malignant. The dataset

¨The class Imbalance Ratios (IR) of PIMA and Breast Cancer data sets are respectively 1.6 and 1.8.

4.4 Performance Measures

For the Active SMOTE based models, by the end of each epoch, the classification model is evaluated on the test data set. Results are recorded in a confusion matrix from which accuracy, $F1 - weigth$, recall, sensitivity, and AUC are calculated. The same measures are computed for the other configurations based on the classic SMOTE and its variants by the end of each learning process. All the results corresponding to one machine learning method (LR, SVC or GB) are illustrated by a single figure in section 5

5 Results and Discussion

Figures 6 and 7 illustrate respectively the results obtained on PIMA and Breast Cancer data sets. The different graphs show on the left line plots the newly proposed methods (that's why they have epochs in the x-axis), while the bar plots from the right are the classical sampling techniques previously described in the subsection 4.2. All the experiments below generate ten new synthetic samples in every epoch with every method. In epoch 18th, the dataset is balanced.

5.1 Results on PIMA data set

The first plot in Figure6 demonstrates clearly how the methods of Active SMOTE, when used with Logistic regression, achieve more outstanding performance in recall than the classical ones and are better than simple iterative SMOTE. The better performance of the minority class comes at the expense of the majority class. That is why there is a slight decrease in Precision, F1-Weighted and Balanced Accuracy. Furthermore, Active SMOTE achieved a similar performance in recall as the classical methods at around iteration 12. That means that with only 120 synthetic samples, Active SMOTE achieved equivalent performance as the classical methods. Since the 13th epoch, Active SMOTE with uncertainty

[4] `http://archive.ics.uci.edu/ml`

margin reaches higher performances of recall and accuracy, largely better than the other variants of balancing.

The second plot in Figure 6 illustrates the results with SVC on PIMA dataset. There is considerable variability with all the methods when using SVC. In fact, there is a decrease in recall as epochs pass. Our intuition is that we need to augment the diversity sampling cluster size to make a more diverse sampling because the model generates the synthetic samples in a single region. In this case, the new methods like the classical ones are not performing well. The undersampling technique achieves the best results in all the metrics. Our intuition is that the hyperspace where the new synthetic samples are generated is wrong. In this example, we can see one of the advantages of Active Learning with SMOTE. If we see that the model is doing a lousy performance as epochs pass, we can ask an expert to verify if our synthetic samples are ok.

The best results on PIMA data are given by the Ensemble method Gradient Boost (Figure6, plot 3). The behavior of these graphs is quite similar to Logistic Regression. There is a considerable increase in recall and not a big decrement in precision with Active SMOTE. As with Logistic Regression, Active SMOTE can achieve similar results as classical methods at around epoch 12. ROC – AUC does not change much as epochs pass.

5.2 Results on Breast Cancer data set

From the first plot in figure 7, we can remark that when the confidence measure margin is used, there is a decrease in precision with the active learning strategy. furthermore, there is a slight variation in the metrics as we generate new synthetic samples, which shows that a large margin already separates the classes.

Similarly, the second plot, illustrating the results with SVC, shows that the worst performance metric with Active Learning is done with margin of confidence as an uncertainty measure. However, the proposed new methods perform better than classical methods in recall. It is interesting to note that SVC is able to deal with the imbalance of this data set and provide satisfactory results with a balancing strategy (case of UNBAL).

The Logistic Regression and SVC algorithms did not show much variance because they are simpler than ensemble techniques. That is why we will see a more significant variance in the results obtained with Gradient Boost. In other words, this is a case of the Bias and variance trade-off. Otherwise, there is a substantial increase in recall from iterations 9 to 10. Indeed, Active SMOTE with uncertainty margin, with or without diversity sampling, gives the best scores in terms of recall, balanced accuracy (score higher than 0.96) and F1-weighted (score higher than 0.98). Those newly generated samples were of great utility for the model. Precision does not decrease when recall increases.

Further discussion The main conclusion that can be deduced from this first experimental study is that it is not necessary to completely balance the data to obtain good learning results. Indeed, Active SMOTE has proven that the learning

algorithm can reach an optimal performance with the first synthetic instances generated during the first epochs. This can avoid, in the case of medical data, filling the training set with synthetic positive cases, affecting the quality of the original data. Thus, it is important to adjust the technique of selecting pivot points for SMOTE to the nature of the available data. For example, diversity sampling gave the best performance with the Diabetes database.

Indeed, many datasets are biased toward a specific gender, race, and socioeconomic background. That bias generally occurs in favor of the most privileged demographic: persons from the wealthiest nations, problems from the wealthiest economies, and other biases resulting from a power imbalance. We believe it is vital to research later if doing diversity sampling with SMOTE shows evidence of increasing the diversity of the people who can benefit from models built from data. Active Learning with SMOTE could also be used to know where to do medical studies to maximize model performance. For example, knowing that the model generates synthetic samples in a particular age group could enforce the decision to study that age group in real life.

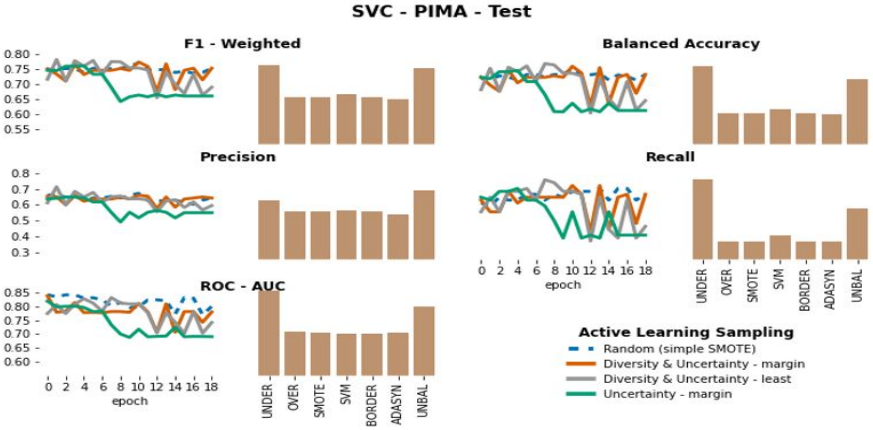
6 Conclusion

We introduced in this paper Active SMOTE, a new dynamic approach to deal with imbalanced medical data. Active SMOTE combines SMOTE with Active Learning using specific sampling techniques based on diversity and/or uncertainty in the positive class.

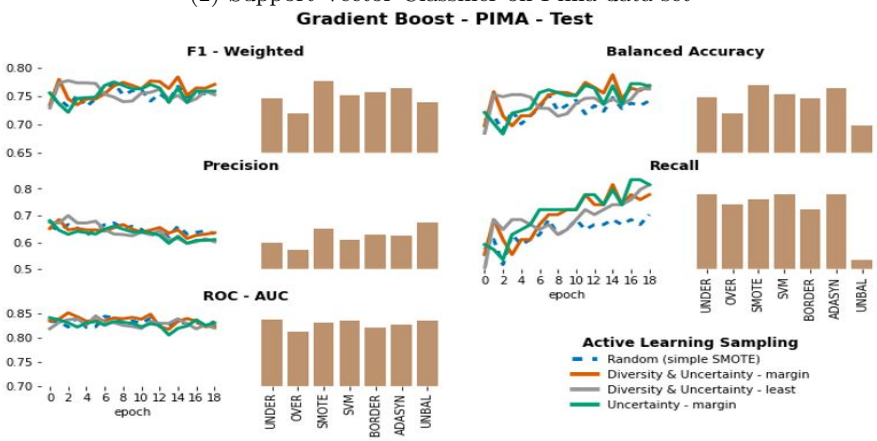
Combined with three machine learning techniques, Logistic Regression, Support Vector classifier and Gradient Boost, and applied to two medical data sets for diabetes diagnosis and breast cancer diagnosis, Active SMOTE has proven its ability to improve the performance of the obtained classifiers in several cases, mainly when applied with an Ensemble Learner. It also demonstrated, thanks to the active learning strategy, that it is not necessary to completely balance the training set to reach high satisfactory results, which is very important in the medical context, since it could reduce the synthetic data corresponding to fictive patients.

The results presented in this paper are obtained with a preliminary experimental study. Further experiments will be done on other data sets with higher imbalance rate and for further medical purpose.

(1) Results of Logistic Regression on Pima data set

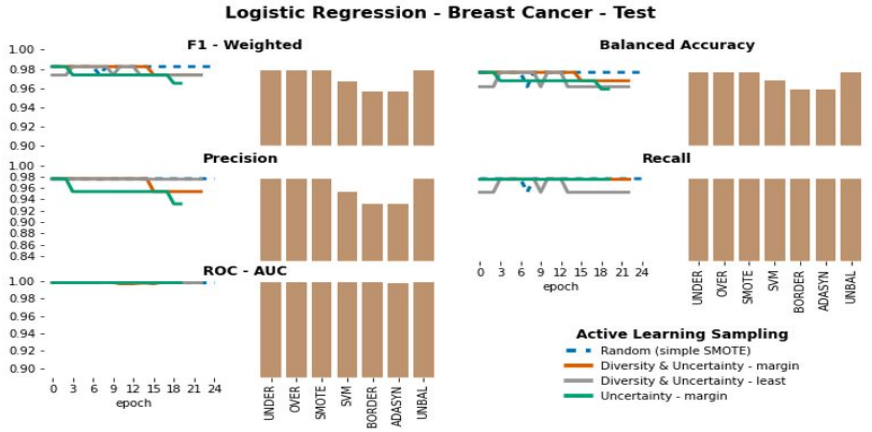


(2) Support Vector Classifier on Pima data set

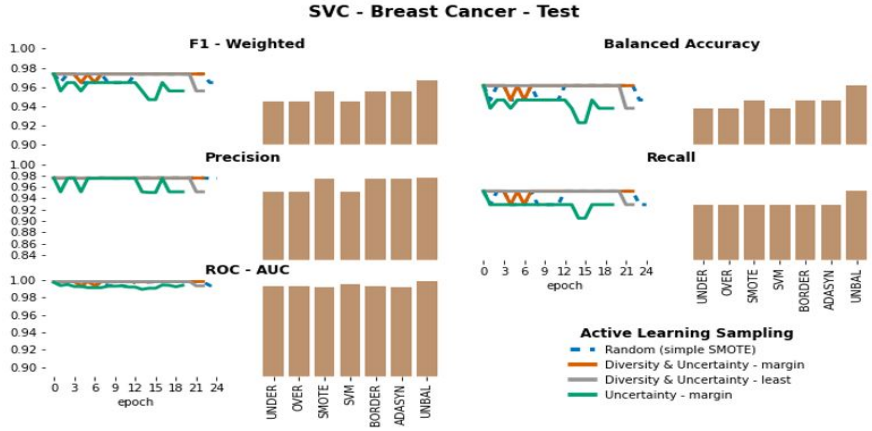


(3) Results of Gradient Boost on Pima data set

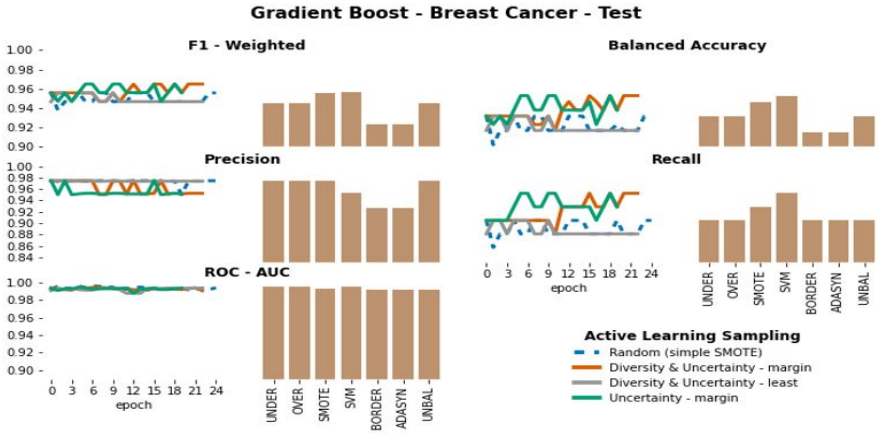
Fig. 6. Active SMOTE on Pima data set



(1) Results of Logistic Regression on Breast Cancer data



Results Support Vector Classifier on Breast Cancer data



Results Gradient of Boosting on Breast Cancer data set

Fig. 7. Active SMOTE: Results on Breast Cancer data set

References

1. Aggarwal, C.C., Kong, X., Gu, Q., Han, J., Philip, S.Y.: Active learning: A survey. In: Data Classification, pp. 599–634. Chapman and Hall (2014)
2. Bach, F.R., Heckerman, D., Horvitz, E.: Considering cost asymmetry in learning classifiers. The Journal of Machine Learning Research **7**, 1713–1741 (2006)
3. Chawla, N.V., Japkowicz, N., Kotcz, A.: Special issue on learning from imbalanced data sets. ACM SIGKDD explorations newsletter **6**(1), 1–6 (2004)
4. Chen, B., Xia, S., Chen, Z., Wang, B., Wang, G.: Rsmote: A self-adaptive robust smote for imbalanced problems with label noise. Information Sciences **553**, 397–428 (2021). https://doi.org/https://doi.org/10.1016/j.ins.2020.10.013
5. Devarriya, D., Gulati, C., Mansharamani, V., Sakalle, A., Bhardwaj, A.: Unbalanced breast cancer data classification using novel fitness functions in genetic programming **140**, 112866. https://doi.org/https://doi.org/10.1016/j.eswa.2019.112866
6. Dua, D., Graff, C.: UCI machine learning repository (2017), `http://archive.ics.uci.edu/ml`
7. Elreedy, D., Atiya, A.F.: A comprehensive analysis of synthetic minority oversampling technique (smote) for handling class imbalance. Information Sciences **505**, 32–64 (2019)
8. Ertekin, S., Huang, J., Giles, C.L.: Active learning for class imbalance problem. In: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. p. 823–824. SIGIR '07, ACM (2007). https://doi.org/10.1145/1277741.1277927
9. Fernández, A., García, S., Galar, M., Prati, R.C., Krawczyk, B., Herrera, F.: Learning from imbalanced data sets, vol. 10. Springer (2018)
10. Fernandez, A., Garcia, S., Herrera, F., Chawla, N.V.: SMOTE for learning from imbalanced data: Progress and challenges, marking the 15-year anniversary **61**, 863–905 (2018). https://doi.org/10.1613/jair.1.11192
11. Hamida, S.B., Benjelloun, G., Hmida, H.: Trends of evolutionary machine learning to address big data mining. In: 5th International Conference, ICIKS, 2021, Proceedings. Lecture Notes in Business Information Processing, vol. 425, pp. 85–99. Springer (2021). https://doi.org/10.1007/978-3-030-85977-0_7
12. Han, H., Wang, W.Y., Mao, B.H.: Borderline-smote: a new over-sampling method in imbalanced data sets learning. In: International conference on intelligent computing. pp. 878–887. Springer (2005)
13. He, H., Bai, Y., Garcia, E.A., Li, S.: 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
14. Hmida, H., Hamida, S.B., Borgi, A., Rukoz, M.: Sampling methods in genetic programming learners from large datasets: A comparative study. In: Advances in Big Data - Proceedings of the 2nd INNS Conference on Big Data, 2016, Thessaloniki, Greece. pp. 50–60 (2016). https://doi.org/10.1007/978-3-319-47898-2_6
15. Le, T., Vo, M.T., Vo, B., Lee, M.Y., Baik, S.W.: A hybrid approach using oversampling technique and cost-sensitive learning for bankruptcy prediction. Complexity **2019** (2019)
16. Li, J., Zhu, Q., Wu, Q., Zhang, Z., Gong, Y., He, Z., Zhu, F.: Smote-nan-de: Addressing the noisy and borderline examples problem in imbalanced classification by natural neighbors and differential evolution. Knowledge-Based Systems **223**, 107056 (2021)

17. Oh, S., Lee, M.S., Zhang, B.T.: Ensemble learning with active example selection for imbalanced biomedical data classification **8**(2), 316–325. https://doi.org/10.1109/TCBB.2010.96, conference Name: IEEE/ACM Transactions on Computational Biology and Bioinformatics
18. Pazzani, M., Merz, C., Murphy, P., Ali, K., Hume, T., Brunk, C.: Reducing misclassification costs. In: Machine Learning Proceedings 1994, pp. 217–225. Elsevier
19. Sáez, J.A., Luengo, J., Stefanowski, J., Herrera, F.: Smote-ipf: Addressing the noisy and borderline examples problem in imbalanced classification by a re-sampling method with filtering. Information Sciences **291**, 184–203 (2015). https://doi.org/https://doi.org/10.1016/j.ins.2014.08.051
20. Xu, Z., Shen, D., Nie, T., Kou, Y., Yin, N., Han, X.: A cluster-based oversampling algorithm combining smote and k-means for imbalanced medical data. Information Sciences **572**, 574–589 (2021)
21. Zhang, J., Wu, X., Shengs, V.S.: Active learning with imbalanced multiple noisy labeling. IEEE Transactions on Cybernetics **45**(5), 1095–1107 (2015). https://doi.org/10.1109/TCYB.2014.2344674