Comparison of Oversampling Algorithms to Classify Imbalanced Data

Andres Cambronero

In applied research, imbalanced datasets are common and often unavoidable. An imbalanced dataset is one in which the number of observations belonging to one class of the response variable greatly exceeds the number of observations in the other class or classes. For example, in medical research, many more pacients are likely to test negative for a particular disease than positive. If such data was collected, it would result in an imbalanced dataset. In these research areas, classification methods can greatly enhance the quality of studies.

Unfortunately, classification methods often perform poorly when trained on imbalanced datasets. Common classification methods minimize the overall error rate; they do not attempt to correctly predict the "rare" class label. In many fields where imbalanced datasets are common, this behavior is problematic because correctly predicting the rare case is often more important than predicting the majority class. As an example, a false negative provides a patient with an incorrect diagnosis that could delay life-saving treatment. By employing standard classifications methods on imbalanced data, researchers inadvertently ignore the costs of false negatives.

To solve this problem, researchers have developed several strategies. A particularly simple method is to oversample the minority class and train the model on this artificially balanced dataset. Such approach changes the prior probabilities imposed on the majority and minority class, which allows for greater separability between classes. A popular oversampling algorithm, Synthetically Minority Oversampling Technique (SMOTE), follows this approach. SMOTE balances the class sizes in the training set by creating artificial data points based on the characteristics of minority class observations. Since the advent of SMOTE, a number of similar preprocessing algortihms have emerged to deal with imbalanced datasets in classification tasks.

This paper compares the performance of a Support Vector Machines that are trained using several oversampling techniques to classify observations in an imbalanced dataset. In particular, the sampling techniques used are Random Oversampling, Random Undersampling, SMOTE, Borderline SMOTE, ADASYN and Safe-Level SMOTE. The results show that, using this particular dataset, SVMs ability to correctly classify minority class observations increases when using SMOTE and its variants compared to training the model on the imbalanced data. However, SVM's ability to classify minority class observations when using SMOTE and its variants is not substantially better than when simply balancing the class observations randomly. While this result might be specific to the dataset used, this result refutes previous findings that suggest that SMOTE and its variants lead to better classification performance than random oversampling.

Literature Review

The literature on imbalanced data and classification methods suggests two general measures to improve the performance of classifiers: sampling and cost-sensitive learning. Breiman et al (2004) investigate the effect of under-sampling from the majority class, over-sampling the minority class, and increasing the cost of misclassifying minority class observations using Random Forest. Using performance metrics appropriate for imbalanced datasets, their results suggest that both sampling approaches led to a larger improvement in performance than cost-sensitive learning.

Van Hulse et. al (2007) delve into the difference between simple and intelligent sampling techniques used to overcome imbalanced data. The authors compare the performance of SVM, K-nearest neighbors, decision tress, logistic regression, and Naive Bayes when trained on data that has been balanced via: random undersampling, random oversampling, one-sided selection, cluster-based oversampling, Wilson's editing, SMOTE, and borderline-SMOTE. Applying these methods over several datasets, the authors find that random oversampling improved the performance of Random Forest and logistic regression. The researchers note that, while a sampling method's performance is dependent on the metric used, the intelligent sampling techniques SMOTE, borline-SMOTE, one-sided selection, and cluster-based oversampling led to inferior performance compared to simple sampling techniques. These results suggest that the performance of learners varies between sampling techniques.

Similarly, He et al (2008) compare the performance of SMOTE with a novel oversampling technique: Adaptive Synthetic Sampling. Using this method, the authors analyze the performance of a tree-based model trained on the original imbalanced data, SMOTE data and ADASYN data. The researchers find that ADASYN provided the best performance in terms of G-Mean over the six imbalanced datasets. The result implies that the classifier achieved a decent balance of correct classification of observations belonging to the majority and minority classes. This method is one variation on oversampling techniques that can arguably outperform SMOTE.

In their work, Han et al (2005) demonstrate that Borderline-SMOTE might improve the performance of classifiers compared to SMOTE in the presence of imbalanced data. Using four datasets with different levels of class imbalance, their experimental results suggest that Borline-SMOTE can outperform SMOTE in several performance metrics. Specifically, over the four datasets, Borderline-SMOTE outperformed SMOTE in the true positive rate and F-value. With these results, the authors conclude that Borderline-SMOTE might be an improvement over existing algorithms.

Similar to the previous articles mentioned, Bunkhumpornpat et al. (2009) propose the use of Safe-Level SMOTE as an improvement on existing SMOTE-variants. Using decision trees, Naïve Bayes, and SVMs, the authors compare their method to SMOTE and Borderline-SMOTE using precision, recall, F-value, and AUC. The authors find that decision tress trained on Safe-Level-SMOTE achieved higher precision and F-value than when trained on SMOTE and Borderline-SMOTE. While the authors indicate that their method did not lead to a consistent improvement over all metrics, all classifiers and all datasets tested, their method did demonstrate some improvements over SMOTE and Borderline-SMOTE.

Methodology

This section briefly describes the workings of each sampling method used in this paper:

Random Oversampling: Using this method, observations from the minority class label in the imbalanced data are randomly sampled with replacement until they represent a desired percent of the balanced data. In this study, the minority class was oversampled until it represented 50 percent of the data.

Random Undersampling: Using this method, observations from the majority class label in the imbalanced data are randomly selected until they represent a desired percent of the balanced data. In this study, the majority class was undersampled until it represented 50 percent of the data.

SMOTE: This method creates artificial data points based on characteristics of existing minority class observations. For each minority class observation, the technique performs k-nearest neighbors. The artificial data points are then introduced along the line segments joining the k-minority class nearest neighbors.

Borderline-SMOTE: Instead of performing SMOTE on all instances of the minority class, this method focuses on creating synthetic observations on the border of the minority decision border. As a result, this method is arguably more computationally efficient than SMOTE.

ADASYN: Similar to Borderline-SMOTE, ADASYN attempts to create synthetic data around particular points in the feature space. In particular, ADASYN generates synthetic observations for minority class observations that are difficult to classify. This method attempts to adaptively shift the decision boundary to those points that are difficult to learn.

Safe-Level SMOTE: As with the oversampling methods above, this method focuses on generating synthetic observations around particular minority class instances. Depending on the number of positive instances surrounding a minority class observations based on its nearest neighbors (safe-level), the method generates a number of number of synthetic data points which are placed on different places along a line.

Using overall accuracy is an inappropriate performance metric for classification methods when in the presence of imbalanced data. If the number of observations in a dataset are distributed in a 9:1 ratio in in a two class response variable, a model that classifies all observations as belonging to the majority class achieves a respectable 90% accuracy rate. As a result, other performance metrics should be considered. Because a model's performance often depends on the metric used to assess it, the models trained on the different oversampling methods are compared on several performance metrics:

- Accuracy= $\frac{TP}{TP+TN}$
- G-mean=$(\frac{TN}{TN+FP} * \frac{TP}{TP+FN})^{\frac{1}{2}}$
- Precision=$\frac{TP}{TP+FP}$
- Recall=$\frac{TP}{TP+FN}$
- F-measure=$\frac{2*Precision*Recall}{Precision+Recall}$

Table 1: Confusion Matrix

	Predicted.Negative	Predicted.Positive
Actual Negative	True Negative (TN)	False Positive (FP)
Actual Positive	False Negative (FN)	True Positive (TP)

These metric definitions were retrieved from Breiman et al (2004). In this case, positive cases refer to the minority class observations and the negative cases refer to majority class observations.

In addition to these metrics, the classifier's performance will be measured using Area under the Receiver Operating Characteristic Curve (AUROC). This metric combines the True Positive rate and False positive rate into a single number. The ROC curve plots true positive observations against false positives for different thresholds, which produces a curve on a plot that ranges from 0 to 1 both axes. A random random predictor would result in a AUROC of 0.5, while a good classifier would achieve a AUROC higher than this value. The closer the AUROC is to 1, the better the classifier performs.

Data Processing

The data used in this report contains information about the proficiency of elementary and middle schools in New York. The final dataset contained 3327 observations. Out of that total, 277 (8%) observations were of class "Proficient" and the remaining were "Not Proficient." The following procedure was followed to merge and clean the dataset into the final stage.

First, the outcome variable "PROF_LEVEL" was created following guidelines of the Every Student Succeeds Act (ESSA) and New York's standard of proficiency. Under ESSA, a school is in need of improvement if English language learners, students receiving special education, racial minorities, or students in poverty consistently underperform. Using the Department of Education's 2015-16 3-8 Assessment Database, each subpopulation of interest within a school received a score of 1 (proficient) or -1 (not proficient) if it's assessment mean achieved the state's standards. Then, each school received a weighted score based on the school's demographics. If the final score was negative, the school was classified as "Not Proficient" and "Proficient" otherwise.

To have a completed dataset, values had to be artificially generated to supplement the information available on the New York State's Department of Education databases, the National Center for Educational Statistics and ACS. These databases contained the following incomplete variables: Suspensions, Average Class Size, Number of Teachers, Percent of Not highly Qualified Teachers, Mean Teacher Score, Enrollment, Expenditure per student, Violent crimes and Property crimes. For these observations, random values were generated from a bounded normal distribution with the mean and variance of the observed values of a school's district. Any remaining missing observations were completed drawing numbers from a normal distribution with the mean and variance of the values observed in the column. With this process, each observation had complete information for each column.

The final dataset contained 26 variables. A full list of variables in the final dataset is provided in the index.

The final dataset contained 3327 observations. Out of that total, 277 (8%) observations were of class "Proficient" and the remaining were "Not Proficient." This dataset was randomly split into a training and test set using a 8:2 ratio. Both subsets retained the same proportions in proficiency level as the original data. Given that the variables were recorded on different scales, all variables except charter status were standardized before analysis.

Results

Below are the confusion matrices showing the SVM predictions of the test data after being trained on each of the datasets described previously. The rows of each matrix show the number of observations belonging to each class, while the columns represent the number of predicted observations for each class. As mentioned previously, the test data contained 56 observed "Proficient" observations and 610 observed "Not Proficient" observations. These matrices are used to compute each of the performance metrics detailed in methodology section.

The confusion matrices show that the imbalanced training set produced the least number of correct predictions (18) for the "Proficient" class, while random-oversampling achieved the most correct observations of this class (55). On the other hand, the SVM trained on the imbalance data achieved the most correct classifications of the "Not Proficient" class, while the SVM trained on the random undersampling achieved the least. This behavior is expected considering that the "Not proficient" label is the majority class. This result is not very useful since the goal of the task is to correctly identify minority class observations.

Table 2: Confusion Matrix Imbalanced Data

	Not.Proficient	Proficient
Not Proficient	604	6
Proficient	38	18

Table 3: Confusion Matrix Random Oversampling

	Not.Proficient	Proficient
Not Proficient	454	156
Proficient	1	55

Table 4: Confusion Matrix Random Undersampling

	Not.Proficient	Proficient
Not Proficient	426	184
Proficient	4	52

Table 5: Confusion Matrix SMOTE

	Not.Proficient	Proficient
Not Proficient	539	71
Proficient	13	43

Table 6: Confusion Matrix Borderline-SMOTE

	Not.Proficient	Proficient
Not Proficient	539	71
Proficient	15	41

Table 7: Confusion Matrix ADASYN

	Not.Proficient	Proficient
Not Proficient	520	90
Proficient	10	46

Table 8: Confusion Matrix Safe-Level SMOTE

	Not.Proficient	Proficient
Not Proficient	556	54
Proficient	20	36

Figure 1 presents the overall accuracy of the SVM model on the test set after being trained on each of the datasets described previously. As expected, the overall accuracy suggests that the classifier performs best when it is trained on the original imbalanced data. Using this training dataset, the SVM correctly classifies 93.4% of all observations. This result is followed by the SVM trained using the data balanced using safe-level SMOTE and SMOTE, which achieved 88.9% and 87.4% accuracy respectively.

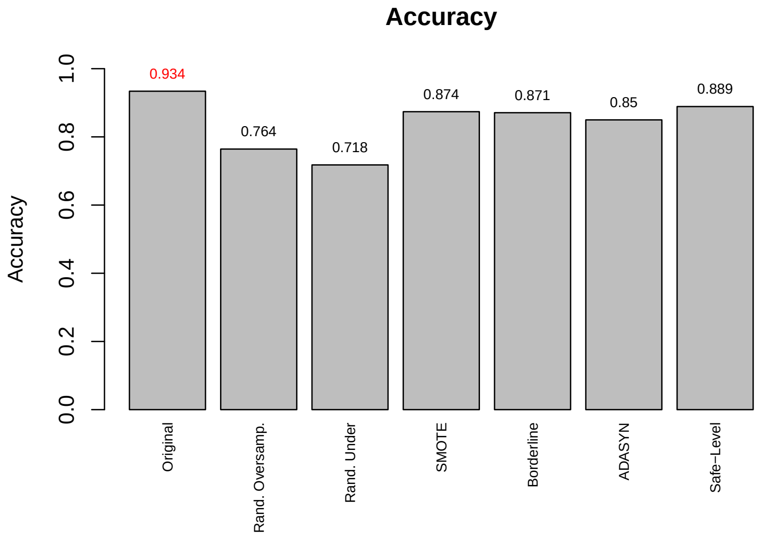

Figure 1: Overall Accuracy of SVM using different sampling methods

While the overall accuracy for the model trained on the original suggest the classifier performs well, its confusion matrix shows that this result is deceiving. As is common when classifying imbalanced data, Table 2 demonstrates that the classifier achieved such high accuracy by predicting most observations as belonging to the majority class ("Not Proficient"). Since the goal of the task is to correctly classify minority class observations, the overall accuracy is a misleading performance metric and the imbalanced training data actually produces poor results.

Figure 2 shows the recall of the SVM trained on each of the datasets being analyzed. The figure indicates that training the SVM on the random oversampling achieves the best recall of 98.4%, followed by random undersampling (92.9%) and ADASYN (82.1%). Looking at Table 3, it is clear that training the SVM on the randomly sampled data produced a high recall, but does so at the expense of misclassifying many observations in the majority “Not Proficient” class. While there is some variation in the SVMs trained on different sampling methods, their recall performance substantially outstrips that of the SVM trained on the original data.

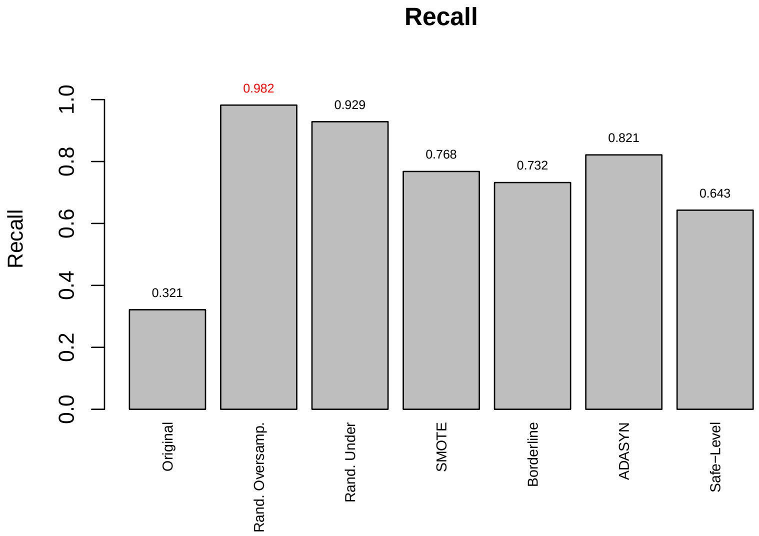

Figure 2: Recall of SVM using different sampling methods

Figure 3 shows the precision of the SVM trained on each of the datasets being analyzed. The figure indicates the SVM trained on the original data achieved the highest precision with 75%. The other sampling methods achieved substantially lower precision levels. Safe-Level SMOTE and SMOTE achieved the second and third highest level at 40% and 37.7% respectively. Table 2 shows that the SVM trained on the imbalanced data achieved such high levels of precision because only six “Not Proficient” observations were misclassified. In an attempt to correctly classify minority observations, the other methods result in a higher number of incorrectly classified observations from the majority class.

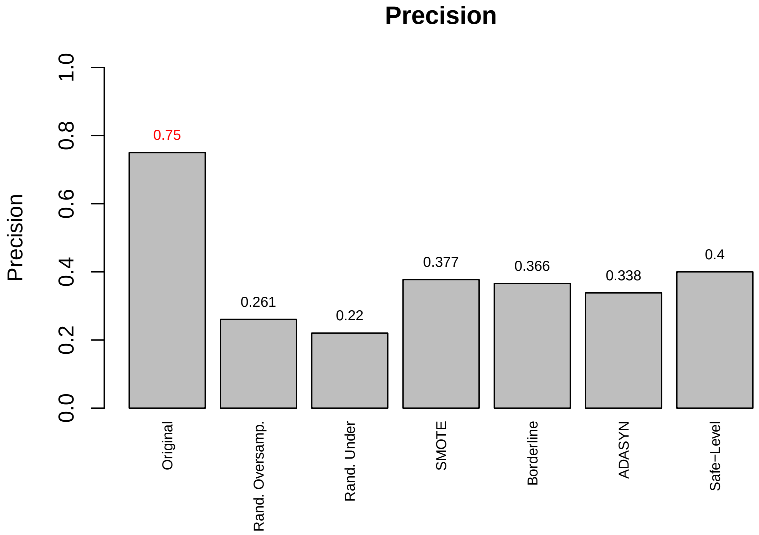

Figure 3: Precision of SVM using different sampling methods

Figure 4 shows the G-mean of the SVM trained on each of the datasets being analyzed. For this performance metric, random oversampling performed the best with a G-mean of 0.855. This sampling method is followed by ADASYN and SMOTE at 0.837 and 0.824 respectively. While the difference in G-mean between the sampling methods is not very large, they all substantially outperform the SVM trained on the imbalanced dataset, which has a G-mean of 0.564. This difference derives from the fact that imbalanced data achieved a very low true positive rate but a high true negative rate, while the other methods achieved some balance between both concepts.

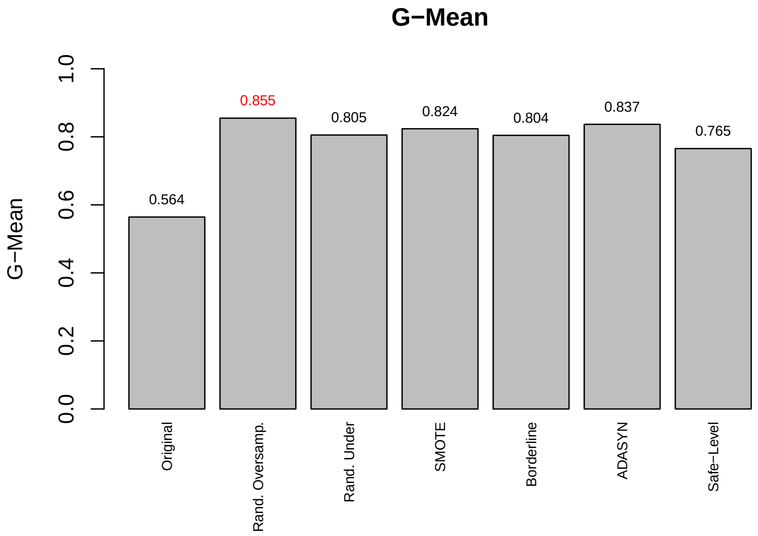

Figure 4: G-Mean of SVM using different sampling methods

Figure 5 shows the F-measure of the SVM trained on the different training datasets. Using this performance metric, the SVM using SMOTE performs best with a measure of 0.506, while random undersampling performs the worst at 0.356. Unlike when comparing other performance metrics, the difference in F-measure between the SVM trained on the original data and the sampling methods is quite small. This result highlights the notion that a classifier's performance depends on the performance metric used to evaluate it.

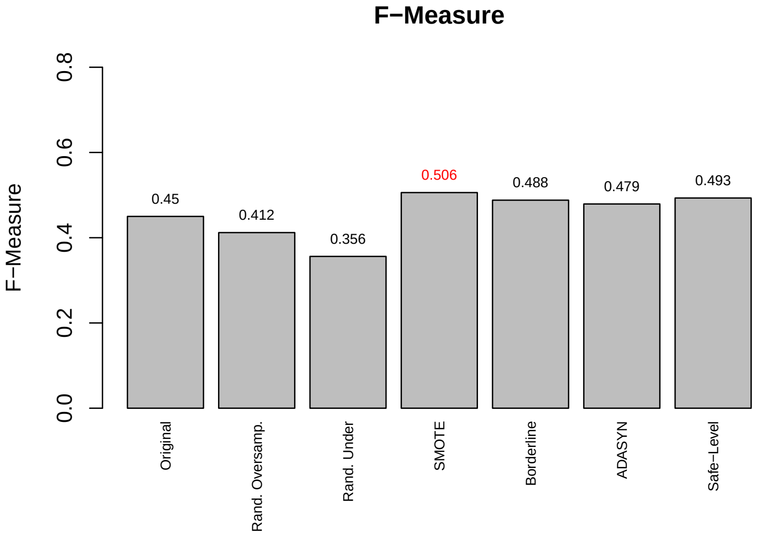

Figure 5: F-measure of SVM using different sampling methods

Figure 6 shows the Area Under the Receiver Operating Characteristic (AUCROC) for each of the SVM trained on the different sampling methods. Random oversampling achieved the highest AUCROC at 0.963, followed by random undersampling (0.963) and the original data (0.899). The sampling method that achieved the lowest AUCROC was SMOTE. Compared to other performance metrics, the difference in the values achieve between sampling methods is not very large. Using this metric, there is no sampling method that stands out as performing best. Despite the similarity in performance, this result does not fall in line with the results reported by previous studies listed in the literature review.

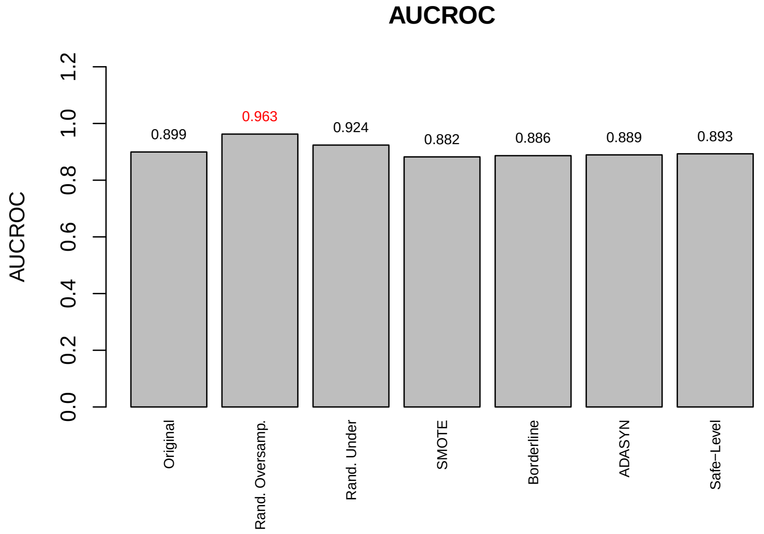

Figure 6: AUCROC of SVM using different oversampling methods

Figure 7 also shows the Area Under the Receiver Operating Characteristic (AUCROC) for each of the SVM trained on the different sampling methods. The plot shows that all sampling methods used performed relatively similarly. While it is difficult to see, Random oversampling achieved the highest AUCROC and SMOTE achieved the lowest.

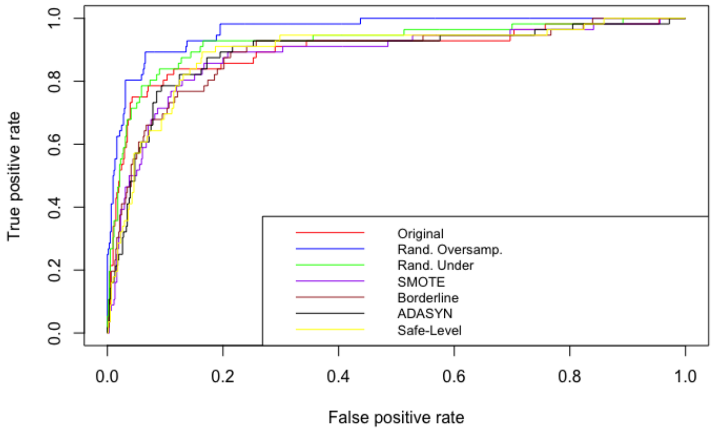

Figure 7: ROC curve

Discussion

Imbalanced data presents a major problem to researchers attempting to use classification methods. It is widely documented that imbalanced training datasets often lead to poor classification results, specially of minority class observations. In fields where imbalanced datasets are common such as medicine and social science, misclassifying minority class observations is often associated with high social and economic costs.

This report compared the performance of a common classification method when trained using sampling methods that studies have suggest are effective at overcoming the challenges of imbalanced data. The results support some of the findings of previous studies and is inconsistent with others. First, as previous studies have reported, using overall accuracy is an inappropriate measure of performance when dealing with imbalanced data. In this case, the SVM trained on the imbalanced data achieved the highest overall accuracy, but did so by misclassifying most minority class observations. Given that classifying minority class observations is often more important, overall accuracy is a poor assessment of a classifier's performance when in presence of imbalanced.

Second, unlike what other studies have found, the SVM trained on the data balanced using random oversampling did not lead to overfitting and actually achieved the highest number of correct predictions of the minority class. In fact, this method only misclassified one observations of the minority class. While it correctly classified most minority class observations, it did not achieve a good balance between classifying majority and minority observations; it misclassified a substantial portion of the majority class observations. In other words, if researchers are interested exclusively in identifying minority class observations and have little interest in majority class observations, random undersampling could an appropriate approach to overcome imbalanced datasets.

The variations of the SMOTE approach that were compared performed quite similarly and did not clearly outperform randomly balancing methods. In none of the performance metrics used did any one particular SMOTE variation stand out. Often, the SMOTE variations performed worse than the random balancing methods. However, unlike the random balancing methods, the SMOTE variations achieved some balance between correctly predicting minority and majority class observations simultaneously. Therefore, if researchers are interested in correctly classifying minority and majority class observations, SMOTE variations might be a more appropriate approach than balancing the dataset via random sampling.

It is important to note that the results of this report should not be generalized. Unlike the studies listed in the literature review, this report only compared the performance of one type of classifier on one single dataset. The performance of the SMOTE variations could outperform the random sampling approach if other classifiers were used or applied to different datasets. For that reason, the results of the report are limited to the particular type of SVM and dataset used.

References

Bunkhumpornpat, Chumphol, Krung Sinapiromsaran, and Chidchanok Lursinsap. “Safe-Level-SMOTE: Safe-Level-Synthetic Minority Over-Sampling TEchnique for Handling the Class Imbalanced Problem.” PAKDD 5476, (2009).

Chawla, N., Bowyer, K., Hall, L., Kegelmeyer, W. “SMOTE: Synthetic Minority Over-Sampling Technique.” Journal of Artificial Intelligence Research 16, 321–357 (2002).

Chen, Chao, Andy Liaw, and Leo Breiman. “Using Random Forest to Learn Imbalanced Data.” (2004).

Han, H., Wang, W., Mao, B. “Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning.” ICIC. (2005).

He, H., Bai, Y., Garcia, E., Li, S. “ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning.” 2008 International Joint Conference on Neural Networks. (2008).

Van Hulse, J., Khoshgoftaar, T. and Napolitino, A. “Experimental Perspectives on Learning from Imbalanced Data.” Proceedings of the 24 th International Conference on Machine Learning, 935-942. (2009).

Index: List of variables

The following is a list of variable abreviations:

```
-PROF_LEVEL: Proficiency level
-PER_FREE_LUNCH: Percent of student body in a school on free lunch
-PER_REDUCED_LUNCH: Percent of student body on reduced lunch
-PER_LEP: Percent of student body that is English Language Learner
-PER_AM_IND: Percent of student body that is Native American
-PER_BLACK: Percent of student body that is Black
-PER_HISP: Percent of student body that is Hispanic
-PER_ASIAN: Percent of student body that is Asian
-PER_WHITE:Percent of student body that White
-PER_Multi: Percent of student body that Multiracial
-PER_SWD: Percent of student body that has a disabiliy
-PER_FEMALE: Percent of student body is female
-PER_ECDIS: Percent of student body that is economically disadvantaged
-PER_SUSPENSIONS: Percent of student body that has received a suspension
-MEAN_CLASS_SIZE: Average class size in grades 3-8
-MEAN_ENROLL: School's average enrollment in 2014-2015
-PER_NOT_HQ: Percent of teachers that are not highly qualified
-STUD_TEACH_RATIO: Ratio of students to teacher
-PER_HS_DEGREE: Percent of people in school district that has a high school degree
-MEAN_INC: Mean income at the school district  level
-EXP_PER_ST: Expenditure per student
-MEAN_TEACH_SCORE: Average teacher assessment score
-CHARTER: Whether school has charter status
-VIOL_CRIME_PER_CITIZEN: Violent crimes per citizen at the county level
-PROP_CRIME_PER_CITIZEN: Property crimes per citizen at the county level
-TOTAL_DIST_POP: Total population at the school district level
```