International Journal of Advances in Engineering and Management (IJAEM)
Volume 2, Issue 12, pp: 53-56 www.ijaem.net ISSN: 2395-5252

Balancing Technique in Medical Data Analysis Using SMOTE Case study on Diabetes Data

PAPINENI NIMISHA
Pondicherry University

Date of Submission: 25-12-2020 Date of Acceptance: 03-01-2021

ABSTRACT: Class imbalance may be a common problem with most medical datasets. Medical datasets are often not balanced in their class labels. A well-balanced dataset is incredibly important for creating an honest prediction model. Most existing classification methods tend to perform poorly on minority class examples when the dataset is extremely imbalanced.Introducing Learning algorithms from imbalanced data has attracted a big amount of interest in recent years. This can be because in universe, imbalanced data exist in many applications, like fault diagnosis, diagnosis, intrusion detection, text classification, financial fraud detection, data stream classification, and soon. In those applications, there are often one or many minority classes possessing only a few samples compared with the opposite classes. And most of your time, the "small" classes are more important than the "large" ones. Thanks to the unbalance of information distribution of imbalanced learning problems, it's often difficult to get good performance for many cases; using traditional classifiers where a balanced distribution of classes is assumed to be an equal misclassification cost for every class is assigned. This can be because they aim to optimize the accuracy without considering the relative distribution of every class. Here I have got considered a case study of diabetic data where well-known over-sampling technique SMOTE is employed and a few under-sampling techniques also are explored. Also, an improved under sampling technique is proposed. Experimental results show that the proposed method displays wonderful performance than prevailing methods. Sampling strategies won't overcome the category imbalance problem by either oversampling or under-sampling. Many researchers have already proposed different methods of over-sampling and under-sampling by which the bulk class sample are often balanced. We examined the popular over-sampling technique SMOTE and a few under sampling techniques over diabetes data. This paper proposes a modified cluster-based under-sampling method that may not only balance the info but can also generate good quality training sets for building classification models. The end result labels of most of the clinical datasets don't seem to be in line with the underlying data. During this paper we examine the performance of over-sampling and under-sampling techniques to balancing diabetic data. If we consider this data set, we observe that diabetic risk relies on whether previous patient's records display diseased or not diseased. Both the situations confound a category imbalance problem. The traditional over-sampling and under-sampling technique might not always be appropriate for such datasets. The proposed method is found to be useful for such datasets where the category labels do not seem to be certain and may also help to beat the category imbalance problem of clinical datasets and for other data domains and also the variable importance in testing a disease to be positive or negative is seen in our analysis.

I. REBALANCING STRATEGIES:

In this context, oversampling and under sampling are the 2 main balancing techniques applied. As a result, traditional classifiers tend to be overwhelmed by the bulk classes and ignore the minority ones, which isn't acceptable in many real applications. While in most cases SMOTE seems beneficial with low-dimensional data, it doesn't attenuate the bias towards the classification within the majority class for many classifiers when data are high-dimensional, and it's less effective than random under sampling. SMOTE is helpful for ANN classifiers for high-dimensional data if the quantity of variables is reduced performing some style of variable selection; we explain why, otherwise, the ANN classification is biased towards the minority class. Furthermore, we show that on high-dimensional data SMOTE doesn't change the class-specific mean values while it decreases the information variability and it introduces correlation between samples. We explain how our findings impact the class-prediction for high-dimensional

International Journal of Advances in Engineering and Management (IJAEM)
Volume 2, Issue 12, pp: 53-56 www.ijaem.net ISSN: 2395-5252

data. In practice, within the high-dimensional setting only ANN classifiers supported the Euclidean distance seem to profit substantially from the utilization of SMOTE, given that variable selection is performed before using SMOTE; the benefit is larger if more neighbours are used. SMOTE for ANN (ARTIFICIAL NEURAL NETWORKS) without variable selection shouldn't be used, because it strongly biases the classification towards the minority class. The objective of sophistication prediction (classification) is to develop a rule supported a gaggle of samples with known class membership (training set), which may be accustomed assign the category membership to new samples. Many alternative classification algorithms (classifiers) exist, and that they are supported the values of the variables (features) measured for every sample. Very often the training and/or test data are class-imbalanced. The quantity of observations belonging to every class isn't the identical. The matter of learning from class-imbalanced data has been receiving a growing attention in many various fields. The presence of class-imbalance has important consequences on the educational process, usually producing classifiers that have poor predictive accuracy for the minority class which tend to classify most new samples within the majority class; during this setting the assessment of the performance of the classifiers is additionally critical. Data are nowadays increasingly often high-dimensional: the quantity of variables is extremely large and greatly exceeds the quantity of samples. Despite the growing number of applications using high-dimensional class-imbalanced data, this problem has been seldom addressed from the methodological point of view. It had been previously shown for several classifiers that the class-imbalance problem is exacerbated when data are high-dimensional the high-dimensionality further increases the bias towards the classification into the bulk class, even when there's no real difference between the classes. The high dimensionality affects each variety of classifier during a different way. A general remark is that giant discrepancies between training data and true population values are more likely to occur within the minority class, which includes a larger sampling variability: therefore, the classifiers are often trained on data that don't represent well the minority class. The high dimensionality contributes to the present problem as extreme values don't seem to be exceptional when thousands of variables are considered. Some of the solutions proposed within the literature to attenuate the class-imbalance problem are effective with high-dimensional data, while others aren't. Generally under-sampling techniques, aimed toward producing a class-balanced training set of smaller size, are helpful, while simple oversampling isn't. The rationale is that in most cases simple oversampling doesn't change the classification rule. Similar results were obtained also for low-dimensional data. The Synthetic Minority Over-Sampling Technique (SMOTE) is an oversampling approach that makes synthetic minority class samples. It potentially performs better than simple oversampling and it's widely used. As an example, SMOTE was used for detecting network intrusions or sentence boundary in speech, for predicting the distribution of species or for detecting diabetes. SMOTE is employed also in bioinformatics for mirnagene prediction, for the identification of the binding specificity of the regulatory proteins and of photoreceptor-enriched genes supported expression data, and for histopathology annotation.

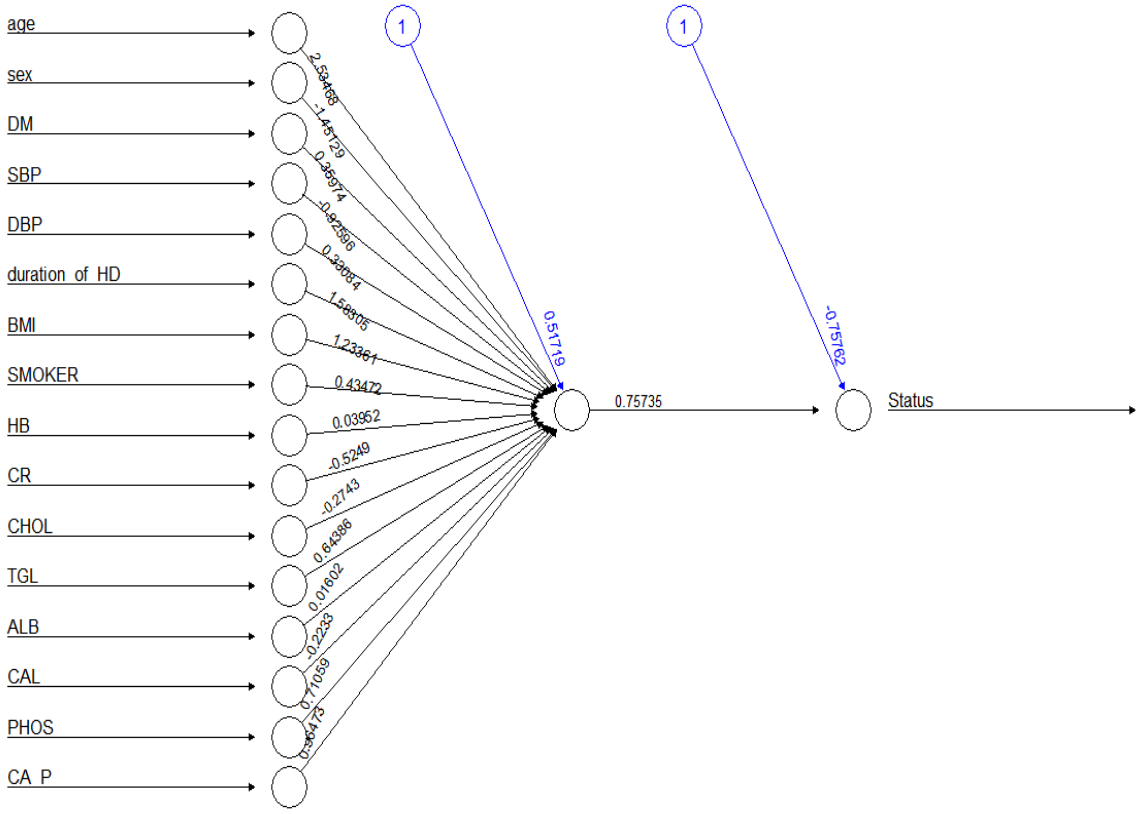
II. INTERPRETATION:

Here firstly we have the summaries of unbalanced data and balanced data separately which interprets that this SMOTE doesn’t make much difference in data’s summaries. This interprets that balancing data is a good choice.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 2, Issue 12, pp: 53-56 www.ijaem.net ISSN: 2395-5252

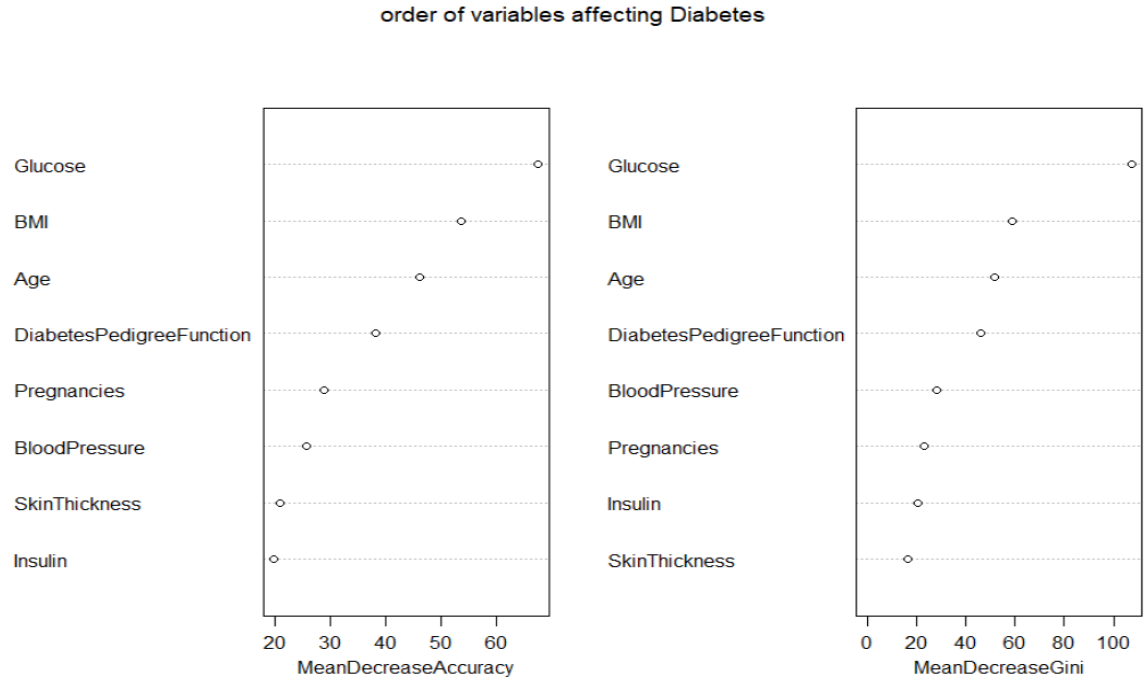
```
> summary(balanced_data_diabetic)
  Pregnancies        Glucose      BloodPressure    SkinThickness      Insulin           BMI
 Min.   : 0.000   Min.   :  0.0   Min.   :  0.00   Min.   : 0.00   Min.   :  0.00   Min.   : 0.00
 1st Qu.: 1.000   1st Qu.:102.0   1st Qu.: 64.00   1st Qu.: 0.00   1st Qu.:  0.00   1st Qu.:28.00
 Median : 3.000   Median :122.0   Median : 72.00   Median :23.00   Median : 22.50   Median :32.80
 Mean   : 4.062   Mean   :125.4   Mean   : 70.34   Mean   :20.66   Mean   : 82.89   Mean   :32.77
 3rd Qu.: 7.000   3rd Qu.:147.0   3rd Qu.: 80.00   3rd Qu.:33.00   3rd Qu.:135.00   3rd Qu.:37.10
 Max.   :15.000   Max.   :199.0   Max.   :122.00   Max.   :99.00   Max.   :846.00   Max.   :59.40
 DiabetesPedigreeFunction      Age        Outcome
 Min.   :0.0780           Min.   :21.00   0:500
 1st Qu.:0.2490           1st Qu.:25.00   1:500
 Median :0.3800           Median :31.00
 Mean   :0.4827           Mean   :34.23
 3rd Qu.:0.6475           3rd Qu.:42.00
 Max.   :2.4200           Max.   :81.00
> summary(diabetes_data)
  Pregnancies        Glucose      BloodPressure    SkinThickness      Insulin           BMI
 Min.   : 0.000   Min.   :  0.0   Min.   :  0.00   Min.   : 0.00   Min.   :  0.0   Min.   : 0.00
 1st Qu.: 1.000   1st Qu.: 99.0   1st Qu.: 62.00   1st Qu.: 0.00   1st Qu.:  0.0   1st Qu.:27.30
 Median : 3.000   Median :117.0   Median : 72.00   Median :23.00   Median : 30.5   Median :32.00
 Mean   : 3.845   Mean   :120.9   Mean   : 69.11   Mean   :20.54   Mean   : 79.8   Mean   :31.99
 3rd Qu.: 6.000   3rd Qu.:140.2   3rd Qu.: 80.00   3rd Qu.:32.00   3rd Qu.:127.2   3rd Qu.:36.60
 Max.   :17.000   Max.   :199.0   Max.   :122.00   Max.   :99.00   Max.   :846.0   Max.   :67.10
 DiabetesPedigreeFunction      Age        Outcome
 Min.   :0.0780           Min.   :21.00   0:500
 1st Qu.:0.2437           1st Qu.:24.00   1:268
 Median :0.3725           Median :29.00
 Mean   :0.4719           Mean   :33.24
 3rd Qu.:0.6262           3rd Qu.:41.00
 Max.   :2.4200           Max.   :81.00
```

Artificial neural netorks applied for this balanced data and the graph below represents the contribution of each variable in having diabetes.



The below graph represents variable importance order in showing whether the patient is confirmed positive for diabetes or negetive for diabetes based on the level of each variable

International Journal of Advances in Engineering and Management (IJAEM)
Volume 2, Issue 12, pp: 53-56 www.ijaem.net ISSN: 2395-5252



III. CONCLUSIONS:

Now recently experimentally observed using low-dimensional data that simple under sampling tends to outperform SMOTE in most situations. This result was further confirmed using SMOTE with SVM as a base classifier, extending the observation also to high-dimensional data.SMOTE with SVM seems beneficial but less effective than simple under sampling for low-dimensional data, while it performs very similarly to uncorrected SVM and generally much worse than under sampling for high-dimensional data. This is an attempt to investigate explicitly the effect of the high-dimensionality on SMOTE, while the performance of SMOTE on high-dimensional data was not thoroughly investigated for classifiers other than SVM. Others evaluated the performance of SMOTE on large data sets, focusing on problems where the number of samples, rather than the number of variables was very large. Several works focused on improving the original SMOTE algorithm, but these modifications were mainly not considered in the high-dimensional context. Random forest technique is also applied in this SMOTE to know which variable is affecting more in having diabetes followed by ANN

REFERENCES:

[1]. A comprehensive data level analysis for cancer diagnosis on imbalanced DATA SARA FOTOUHI [1], SHAHROKH ASADI [2], MICHAEL W KATTAN [3]
[2]. Classification: A case of look-alike sound-alike mix-up incident detection YANG ZHAO,[1] ZOIE SHUI-YEE WONG
[3]. Medical imbalanced data classification Sara Belarouci[a)], Mohammed Amine Chikh
[4]. Addressing the Class Imbalance Problem in Medical Datasets M. Mostafizur Rahman and D. N. Davis