JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

BALANCING THE IMBALANCED DATASET USING SMOTE-ENN

Dr.D.Gandhimathi, Assistant professor, The American College, Madurai
Dr.A.JohnSanjeevKumar, Head of the Department, Department of Data Science, The American college, Madurai
Mr.K.Mukunthan, III BSc DSC, The American College, Madurai

ABSTRACT

The process of balancing the imbalanced data is that there are many methods for balancing the data but only some of them produce a more accurate result. Here the churn dataset of a company is taken as the imbalanced data. Clearly while pre-processing the data we came to know that there are many missing values, and the missing value is replaced using regressing because it would be not appropriate to move on to further process with missing values. After filling out the missing values in the data the value in the data is converted into numerical form i.e., in zeros and ones using the user defined function that we have created. It is for only two value columns which have values like YES or NO as 1 and 0. After converting the value, check out the unique value in the other columns. Categorizing such types of columns. And splitting the columns according to their unique values and assigning the value as zeros and ones. After the cleaning process is done. The SMOTE-ENN method is used to balance the data.

KEYWORDS – UP Sampling, DOWN Sampling, SMOTE (Synthetic Minority Oversampling), ENN (Edited Nearest Neighbor)

1. INTRODUCTION

Classification problems are quite common in the machine learning world. As we know in the classification problem we try to predict the class label by studying the input data or predictor where the target or output variable is a categorical variable in nature. In classification problems, there are instances where one of the target class labels' numbers of observation is significantly lower than other class labels. This type of dataset is called an imbalanced class dataset which is very common in practical classification scenarios. Any usual approach to solving this kind of machine learning problem often yields inappropriate results. Imbalanced data refers to those types of datasets where the target class has an uneven distribution of

observations, i.e one class label has a very high number of observations and the other has a very low number of observations.

The process of balancing the imbalanced data is that there are many methods for balancing the data but only some of them produce a more accurate result. Machine learning techniques for class imbalanced data Addressing class imbalance with traditional machine learning techniques has been studied extensively over the last two decades. The bias towards the majority class can be alleviated by altering the training data to decrease imbalance, or by modifying the model's underlying learning or decision process to increase sensitivity towards the minority group. As such, methods for handling class imbalance are grouped into data-level techniques, algorithm-level methods, and hybrid approaches. This section summarizes some of the more popular traditional machine learning methods for handling class imbalance.

2. LITERATURE REVIEW

As per Literature study, the imbalanced dataset is taken and then it is balanced by the UP-Sampling and DOWN-Sampling method. Due to the imbalanced dataset that we have taken there are two classes i.e., Majority and the Minority classes. The UP-Sampling method is that, where the minority class data is filled up till it matches with the majority class. The DOWN-Sampling is that the majority class is reduced till it matches with the minority class. The problems in these methods are: In UP-Sampling for the minority class duplicate records are created which are fake, In DOWN-Sampling for the majority class the records are removed in accordance with the minority class. This came to the conclusion that it is not precise and accurate. Data-level methods.

Data-level methods for addressing class imbalance include oversampling and under-sampling. These techniques modify the training distributions in order to decrease the level of imbalance or reduce noise, e.g. mislabelled samples or anomalies. In their simplest forms, random under-sampling (RUS) discards random samples from the majority group, while random over-sampling (ROS) duplicates random samples from the minority group. Under-sampling voluntarily discards data, reducing the total amount of information the model has to learn from. Over-sampling will cause an increased training time due to the increased size of the training set, and has also been shown to cause over-fitting. Over-fitting, characterized by high variance, occurs when a model fits too closely to the training data and is then unable to generalize to new data. A variety of intelligent sampling methods have been developed in an attempt to balance these trade-offs.

3. PROBLEM DEFINITION

The Problem definition is that while we are balancing the dataset with Up-Sampling or Down-Sampling methods, the data are balanced without accuracy and precision. In the Up-Sampling method the minority class is balanced with the majority class while creating the duplicate value for the minority class until it balances with the majority class. In the Down-Sampling method the majority class is reduced until it is equal with the minority class. The data are removed in the down-sampling method. In both of the above methods used in balancing the data the accuracy and precision is low after balancing the

data. The Proposed problem is proceeded by four modules which are required to balance the imbalanced data set. They are,

- Data Collection
- Data Processing
- Data Balancing
- Data Visualization

4. PROPOSED BALANCING TECHNIQUE

The Proposed Balancing technique is a combination of SMOTE (Synthetic Minority Oversampling TechniquE), which is the method of dealing with imbalanced data, and ENN (Edited Nearest Neighbor), use to find the K-Nearest Neighbor of each observation first, then checking whether the majority class from the observation's K-Nearest neighbor is the same as the observation's class or not. By combining both the methods the imbalanced dataset is balanced with precision and accuracy. This helps in further process in prediction with more accuracy and precision. Intelligent under-sampling methods aim to preserve valuable information for learning. Several Near-Miss algorithms use a K-nearest neighbors (K-NN) classifier to select majority samples for removal based on their distance from minority samples.

One-sided selection was proposed by Kubat and Matwin as a method for removing noisy and redundant samples from the majority class as they are discovered through a 1-NN rule and Tomek links, a K-NN rule that removes misclassified samples from the training set, to remove majority samples from class boundaries. A number of informed over-sampling techniques have also been developed to strengthen class boundaries, reduce over-fitting, and improve discrimination. Chawla et al. Introduced the Synthetic Minority Over-Sampling TechniquE (SMOTE), a method that produces artificial minority samples by interpolating between existing minority samples and their nearest minority neighbors. Several variants to SMOTE, e.g. Borderline-SMOTE and Safe-Level-SMOTE, improve upon the original algorithm by also taking majority class neighbors into consideration. Borderline-SMOTE limits oversampling to the samples near class borders, while Safe-Level-SMOTE defines safe regions to prevent over-sampling in overlapping or noise regions. The above two methods KNN and SMOTE are combined and used for Balancing the data.

5. METHODOLOGY

Data Balancing is one of the techniques in machine learning. It is used to balance the imbalanced data set. Because when we use imbalanced data for further analysis like prediction it would produce the result with low accuracy and precision. So, the process of Data Balancing is applied in the data set to form a balanced data to give more accuracy and precision. At first the imbalanced data is imported and the pre-processing process is started. While pre-processing the missing values in the column are analysed and it is filled with respective methods. Each column has varied data that is to be changed as zeros and one for further processing. If there are more than two types of records in the same column, then the column is splitted and the values are stored as zeros and ones and after all the pre-processing work is finished. The data balancing technique is applied to the data to balance the imbalanced data. The SMOTE-ENN

method is used to balance the data. After balancing the data, it can be used for further processes like prediction which gives us results with more accuracy and precision.

5.1 DATASET DESCRIPTION

The data set that I have taken is the Telco Customer Churn which is IBM sample data sets. The data set includes information about:

- The Collected Data set has 7004 instances and 21 columns.
- Customers who left within the last month – the column is called Churn
- Services that each customer has signed up for – phone, multiple lines, internet, online security, online backup, device protection, tech support, and streaming TV and movies
- Customer account information – how long they've been a customer, contract, payment method, paperless billing, monthly charges, and total charges
- Demographic info about customers – gender, age range, and if they have partners and dependents.

5.2 DATA PREPROCESSING

At first the imbalanced data is imported and the pre-processing process is started. While pre-processing the missing values in the column are analyzed and it is filled with respective methods. Each column has varied data that is to be changed as zeros and one for further processing. If there are more than two types of records in the same column, then the column is splitted and the values are stored as zeros and ones. After confirming that every column has only the two values that are zeros and one the balancing technique is applied to balance the data.

5.2.1 DATA CLEANING

Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset. When combining multiple data sources, there are many opportunities for data to be duplicated or mislabelled. If data is incorrect, outcomes and algorithms are unreliable, even though they may look correct. There is no one absolute way to prescribe the exact steps in the data cleaning process because the processes will vary from dataset to dataset. But it is crucial to establish a template for our data cleaning process so we know what we are doing the right way every time.

5.2.2 DATA TRANSFORMATION

Data transformation is a technique used to convert the raw data into a suitable format that efficiently eases data mining and retrieves strategic information. Data transformation includes data cleaning techniques and a data reduction technique to convert the data into the appropriate form. Data transformation is an essential data preprocessing technique that must be performed on the data before data mining to provide patterns that are easier to understand. Data transformation changes the format, structure, or values of the data and converts them into clean, usable data. Data may be transformed at two stages of the data pipeline for data analytics projects. Organizations that use on-premises data warehouses generally use an ETL (extract, transform, and load) process, in which data transformation is the middle step.

5.2.3 DATA REDUCTION

Dimensionality reduction is a machine learning (ML) or statistical technique of reducing the amount of random variables in a problem by obtaining a set of principal variables. This process can be carried out using a number of methods that simplify the modelling of complex problems, eliminate redundancy and reduce the possibility of the model overfitting and thereby including results that do not belong. The process of dimensionality reduction is divided into two components, feature selection and feature extraction. In feature selection, smaller subsets of features are chosen from a set of many dimensional data to represent the model by filtering, wrapping or embedding. Feature extraction reduces the number of dimensions in a dataset in order to model variables and perform component analysis

5.3 DATA BALANCING

This method combines the SMOTE ability to generate synthetic examples for minority class and ENN ability to delete some observations from both classes that are identified as having different classes between the observation's class and its K-nearest neighbor majority class. The process of SMOTE-ENN can be explained as follows.

1. (Start of SMOTE) Choose random data from the minority class.
2. Calculate the distance between the random data and its k nearest neighbors.
3. Multiply the difference with a random number between 0 and 1, then add the result to the minority class as a synthetic sample.
4. Repeat step number 2–3 until the desired proportion of minority class is met.
5. Determine K, as the number of nearest neighbors. If not determined, then K=3.
6. Find the K-nearest neighbor of the observation among the other observations in the dataset, then return the majority class from the K-nearest neighbor.
7. If the class of the observation and the majority class from the observation's K-nearest neighbor is different, then the observation and its K-nearest neighbor are deleted from the dataset.

Repeat steps 2 and 3 until the desired proportion of each class is fulfilled. (End of ENN)

5.4 DATA VISUALIZATION

Data visualization is the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data. Additionally, it provides an excellent way for employees or business owners to present data to non-technical audiences without confusion. In the world of Big Data, data visualization tools and technologies are essential to analyze massive amounts of information and make data-driven decisions. Something as simple as presenting data in graphic format may seem to have no downsides. But sometimes data can be misrepresented or misinterpreted when placed in the wrong style of data visualization. When choosing to create a data visualization, it's best to keep both the advantages and disadvantages in mind. It is used to visualize the results of accuracy, precision and many others to compare which method is adaptable for

visualization. The Bar Graph for SMOTE – ENN is discussed in fig 5.1

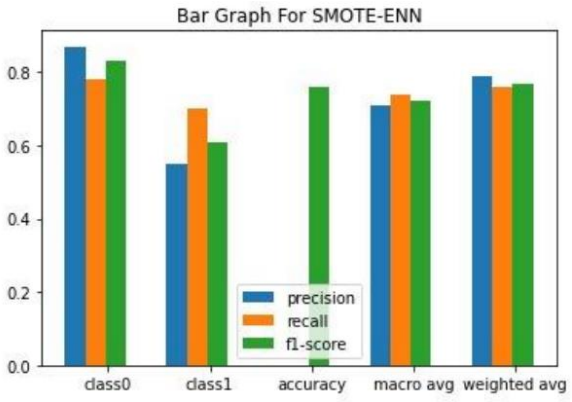

Fig. 5.1 Bar Graph For SMOTE – ENN

6.1 RESULT & DISCUSSION

By applying SMOTE-ENN method other than UP-Sampling or DOWN-Sampling, the accuracy, precision, recall and f1-Score is better than the other two methods. Result of SMOTE – ENN algorithm is discussed in fig 6.1

accuracy_score : 0.7605288932419196

precision_score : 0.5472496473906912

recall_score : 0.697841726618705

f1_score : 0.6134387351778656

	precision	recall	F1 - score
Class 0	0.87	0.78	0.83
Class 1	0.55	0.70	0.61
Accuracy	-	-	0.76
Macro avg	0.71	0.74	0.72
Weighted avg	0.79	0.76	0.77

Fig. 6.1 Results of SMOTE - ENN

By comparing with the above three methods we came to know that the UP-Sampling and DOWN-Sampling are lesser in accuracy, precision when compared to SMOTE-ENN. So we came to the conclusion that it is wise to use the SMOTE-ENN method for balancing the unbalanced data and using it for further improvement or process.

7. CONCLUSION & FUTURE ENHANCEMENT

To know that the overall performance of ML models built on imbalanced datasets, will be constrained by its ability to predict rare and minority points. Identifying and resolving the imbalance of those points is crucial to the quality and performance of the generated models. Balancing the imbalance data is very important in ML in order to achieve the right accuracy. It is not 99% accuracy of the model that matters but the right accuracy of the model 76% matters. Here I selected supervised learning method that is the labelled data for data balancing in order increase the accuracy in terms of prediction result. In future the unlabelled data can also be balanced in order to increase the precision of the data for the further process.

8. REFERENCE

1. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, et al., "Scikit-learn: Machine learning in Python", Journal of Machine Learning Research, vol. 12, pp. 2825-2830, 2011.

2. Z. John Lu, "The elements of statistical learning: data mining inference and prediction", Journal of the Royal Statistical Society: Series A (Statistics in Society), vol. 173, no. 3, pp. 693-694, 2010.

3. I.D. Mienye, Y. Sun and Z. Wang, "An improved ensemble learning approach for the prediction of heart disease risk", Informatics in Medicine Unlocked, vol. 20, pp. 100402, 2020.

4. N.V. Chawla, K.W. Bowyer, L.O. Hall and W.P. Kegelmeyer, "SMOTE: synthetic minority

over-sampling technique", Journal of artificial intelligence research, vol. 16, pp. 321-357, 2002.

5. G.E. Batista, A.L. Bazzan and M.C. Monard, "Balancing Training Data for Automated Annotation of Keywords: a Case Study", WOB, pp. 10-18, 2003.

6. Foster Provost, "Machine Learning from Imbalanced Data Sets 101," the AAAI'2000 Workshop on Imbalanced Data Sets.

7. Nitesh V Chawla, Bowyer Kevin W, Hall Lawrence O, et al. (2002). "Smote: Synthetic Minority Over-Sampling Technique". Journal of Artificial Intelligence Research, Vol. 16, No. 3, pp.321-357.

8. M. Buckland and F. Gey, "The relationship between recall and precision", Journal of the American Society for Information Science, vol. 45, no. 1, pp. 12-19, 1994.

9. Y. Sasaki, "The truth of the F-measure", Teach Tutor Mater, vol. 1, no. 5, pp. 1-5, 2007.