ENHANCEMENT OF PREDICTION USING SYNTHETIC DATA IN HEALTHCARE SYSTEM

S. ASLAM SHAREEF[1], C. THANUJA[2], K. PRAVALLIKA[3], D. RAMYA SAI[4], D. HASEENA[5]
[1]Guide Assistant Professor, [2,3,4,5] U.G.Scholar
[1, 2,3,4,5] Computer Science and Engineering
[1, 2,3,4,5] Ravindra College of Engineering for Women
Email: [1]aslamcse@recw.ac.in, [2] cheerlathanuja14@gmail.com,[3] pravallika2014.k@gmail.com ,
[4]dramya2000@gmail.com, [5]dudekulahaseena66@gmail.com

ABSTRACT

Imbalanced data relates to classification jobs in which the classes are not equally represented. Majority classes are those that make up a considerable share of the data set. Minority classes are those who make up a lesser percentage of the population. Imbalanced datasets are those in which the amount of observations for each class in a classification dataset is different. This imbalance can lead to erroneous outcomes. Because most machine learning algorithms for classification were created with the assumption of an equal number of samples for each class, imbalanced classifications provide a difficulty for predictive modelling.

Despite the fact that many machine learning algorithms have demonstrated considerable success in a variety of real-world applications, the problem of learning from unbalanced data has yet to be solved. Imbalanced learning is a term used to describe learning from unbalanced data.

The following are the major issues with imbalanced learning:

- When a dataset contains underrepresented data, the class distribution begins to bias.
- Learning from such data necessitates new understandings, new approaches, new principles, and new tools to transform data due to the a cost-effective solution to your business problem. In the worst-case scenario, it could result in full trash with no residues to be reused.

Synthetic data is information that is generated artificially rather than through real-world events. Synthetic data is created to meet specific requirements or situations that aren't present in the actual data. Synthetic data is used to safeguard a data set's privacy and confidentiality.

INTRODUCTION

Imbalanced Data Distribution is a phrase used frequently in Machine Learning and Data Science to describe when observations in one class are significantly higher or lower than observations in other classes. Machine Learning algorithms do not address class distribution because their goal is to improve accuracy by minimizing error. This issue is common in areas such as fraud detection, anomaly detection, and facial recognition, among others.

If the classes are not roughly equally represented, the dataset is unbalanced. In fraud detection, imbalances on the order of 100 to 1 are common, and imbalances of up to 100,000 to 1 have been observed in other applications (Provost & Fawcett, 2001). There have been attempts to deal with imbalanced datasets in domains such as detecting oil spills in satellite images (Fawcett & Provost, 1996),

telecommunications management (Ezawa, Singh, & Norton, 1996), text classification (Lewis & Catlett, 1994; Dumais, Platt, Heckerman, & Sahami, 1998; Mladeni'c & Grobelnik, 1999; Lewis & Ringuette, 1994 (Kubat, Holte, & Matwin, 1998).

Predictive accuracy is commonly used to assess the performance of machine learning systems. This is not acceptable, however, when the data is unbalanced and/or the prices of different errors differ significantly. Take the classification of pixels in mammography pictures as potentially malignant as an example (Woods, Doss, Bowyer, Solka, Priebe, & Kegelmeyer, 1993). In a typical mammography dataset, 98 percent of the pixels are normal, whereas only 2% are problematic. A simple default technique of predicting the majority class would result in a 98 percent predicted accuracy. However, the nature of the application necessitates a high rate of correct detection in the minority class while allowing for a low error rate in the majority class. In such cases, simple prediction accuracy is clearly insufficient.

The toy dataset contains 9,900 samples from class 0 and only 100 samples from class 1, resulting in a ratio of 1:100. Assume you're using the above dataset to train your model without accounting for the distribution. When dealing with imbalanced datasets, the most common issue is that a model becomes biassed towards the dominant class. As a result, assigning the label 0 to every single sample is not difficult for a model to attain 99 percent accuracy. However, it is critical to employ a variety of indicators that might provide you with additional information. Our model would be absolutely useless if we were trying to classify spam email. Of course, obtaining more data is always preferable; nevertheless, this may be quite difficult.

There are various techniques of dealing with this problem. Undersampling and Oversampling are two of the most popular and straightforward examples. Oversampling approaches are recommended over Undersampling strategies in most circumstances since eliminating data may result in the loss of crucial features. However, random oversampling may result in overfitting, which is a different issue. It is also possible to combine the two and obtain data that is generally balanced.

The problem of class imbalance has been addressed in two ways by the machine learning community. One option is to give different charges to different training scenarios (Pazzani, Merz, Murphy, Ali, Hume, & Brunk, 1994; Domingos, 1999). Re-sampling the original dataset, either by over-sampling the minority class and/or under-sampling the majority class, is the alternative option (Kubat & Matwin, 1997; Japkowicz, 2000; Lewis & Catlett, 1994; Ling & Li, 1998). Our method (Chawla, Bowyer, Hall, & Kegelmeyer, 2000) combines majority-class under-sampling with a unique type of minority-class over-sampling. Experiments with different datasets using the decision tree classifier (Quinlan, 1992), Ripper (Cohen, 1995b), and a Naive Bayes Classifier reveal that our method outperforms previous re-sampling, altering loss ratio, and class priors methods.

BACKGROUND

The goal of class prediction (classification) is to create a rule that can be used to assign class membership to fresh samples based on a collection of samples with known class membership (training set). There are a variety of classification algorithms (classifiers) available, all of which are based on the values of the variables (features) measured for each sample [1].

The training and/or test data are frequently class-imbalanced, meaning that the number of observations in each class is not equal. Many diverse domains are paying increasing attention to the problem of learning from class-imbalanced data [2]. The presence of class-imbalance has significant implications for the learning process, typically resulting in classifiers with low predictive accuracy for the minority class and a proclivity to classify most new samples in the majority class; in this setting, classifier performance evaluation is also critical [3].

Data is becoming increasingly multidimensional, with a huge number of variables far outnumbering the number of samples. High-throughput technologies, for example, are popular in the biomedical field, where it is possible to assess the expression of all known genes (>20,000) at the same time, but the number of subjects in the study is rarely more than a few hundreds. Many articles attempted to construct classification rules based on class-imbalanced high-dimensional gene expression data (see for example [4-6]).

Despite the increasing number of applications that use high-dimensional class-imbalanced data, this issue has received little attention from a methodological standpoint [2]. The class-imbalance problem is exacerbated for many classifiers when data is high-dimensional [7]: the high-dimensionality increases the bias towards categorization into the majority class, even when there is no substantial difference between the classes. Each type of classifier is affected differently by high-dimensionality. Large disparities between training data and genuine population values are more likely to occur in the minority class, which has higher sampling variability: as a result, classifiers are frequently trained on data that do not accurately reflect the minority class. The difficulty is exacerbated by the large dimensionality of the data, as extreme values are not uncommon when hundreds of variables are included.

Some of the solutions to the class-imbalance problem described in the literature are effective with high-dimensional data, while others are not. Undersampling strategies aiming at producing a smaller, class-balanced training set are often beneficial, but simple oversampling is not [7]. The reason for this is that, in the vast majority of circumstances, simple oversampling has no effect on the categorization rule. Low-dimensional data yielded similar results as well [8].

The Synthetic Minority Over-Sampling TEchnique (SMOTE [9]) is an oversampling technique for producing synthetic minority class samples. It has the potential to outperform basic oversampling and is commonly utilized. SMOTE has been used to detect network intrusions [10], sentence

boundaries in speech [11], species distribution prediction [12], and breast cancer detection [13]. SMOTE is also utilised in bioinformatics for mi-rna gene prediction [14, 15], regulatory protein binding specificity [16], photoreceptor-enriched gene identification based on expression data [17], and histopathological annotation [18].

However, utilizing low-dimensional data, it was recently discovered that simple undersampling outperforms SMOTE in the majority of cases [8]. This result was confirmed using SMOTE with SVM as a base classifier [19], which extended the observation to high-dimensional data: SMOTE with SVM appears to be beneficial but less effective than simple undersampling for low-dimensional data, while it performs very similarly to uncorrected SVM and generally performs much worse than undersampling for high-dimensional data. This was the first attempt, to our knowledge, to explore explicitly the effect of high-dimensionality on SMOTE, despite the fact that the performance of SMOTE on high-dimensional data has not been properly investigated for classifiers other than SVM. Others assessed SMOTE's performance on big data sets, focusing on issues where the number of samples, rather than the number of variables, was extremely large [20, 21]. A number of studies have attempted to improve the original SMOTE algorithm [17, 22-24], but these improvements have mostly been overlooked in the high-dimensional setting.

Imbalanced Data in Healthcare

The imbalance trait found in many real-world healthcare datasets makes classification difficult. Cancer diagnostics, patient safety informatics, and disease risk prediction are all affected by the unbalanced classification problem in the healthcare domain, where data is often severely skewed due to individual individuality and diversity. Most typical classifiers, such as logistic regression and the support vector machine, make the implicit assumption that both classes are equally prevalent. Furthermore, these strategies are intended to improve overall categorization accuracy. As a result, they support the majority, leading to a lack of compassion for minorities. This notion is exemplified in Figure 1, which depicts a synthetic case with a majority and minority class. The solid line () represents the best separator in the underlying distribution, whereas the dotted line () represents the best max-margin loss-minimizing separator constructed over the instances. The induced separator is clearly skewed toward the minority class in this example.

LITERATURE SURVEY

Health care is widely viewed as a critical determinant in fostering people's overall physical, mental, and social well-being, and when well-managed, it may contribute significantly to acountry's economy, development,and industrialization. Healthcare is another industry that evolves with the times. Machine learning algorithms in healthcare have a lot of potential because of the volume of data gathered for each patient.

Machine learning (ML) is a subset of artificial intelligence technology in which algorithms scan massive data sets to find patterns, learn from them, and perform jobs without being told how to solve the problem. The widespread availability of strong hardware and cloud computing has led to a wider acceptance of machine learning in several sectors of human life, ranging from social media recommendations to factory process automation.

1. LOGISTIC REGRESSION

Statistics and data scientists use logistic regression to classify individuals, products, entities, and other things. It's a type of data analysis that creates a binary classification based on one or more independent variables. As a result, it generates two distinct classes (Yes or No, 1 or 0, etc.).

For example, a binary classification could be used to determine if a medical claim is fraudulent or not, or whether or not a patient has diabetes.

For binary classification, logistic regression is an excellent technique. It differs from a lot of other approaches for estimating continuous variables or distributions. This statistical method can be used to determine whether or not a person is more likely to develop cancer as a result of environmental factors such as closeness to a highway, smoking habits, and so on. For a long time, this strategy has been employed successfully in the medical, financial, and insurance industries. It takes time to figure out when to employ which algorithm. However, as a data scientist encounters more difficulties, they will be able to determine whether to utilise logistic regression or decision trees more quickly.

Using logistic regression, healthcare organisations can more precisely target at-risk patients who need a more targeted behavioural health plan to assist them change their everyday health habits. As a result, people will have better health and hospitals would have lower costs.

There are three different types of logistic regression models based on categorical responses.

- **Binary logistic regression**: In this approach, the response or dependent variable is dichotomous in nature—i.e. it has only two possible outcomes (e.g. 0 or 1). Some popular examples of its use include predicting if an e-mail is spam or not spam or if a tumor is malignant or not malignant. Within logistic regression, this is the most commonly used approach, and more generally, it is one of the most common classifiers for binary classification.

- **Multinomial logistic regression**:

The dependent variable in this sort of logistic regression model has three or more possible outcomes, but the order of these values is not specified. To sell films more efficiently, movie studios, for example, aim to anticipate what kind of film a moviegoer is likely to see. A multinomial logistic regression model can assist the studio figure out how much an individual's age, gender, and dating status influence the type of film they prefer. The company can then target a specific movie's advertising campaign at a group of people who are likely to see it.

- **Ordinal logistic regression**:

When the response variable includes three or more alternative outcomes, but these values have a predetermined sequence, this sort of logistic regression model is used. Ordinal responses include grading systems ranging from A to F and rating scales ranging from 1 to 5.

2. RANDOM FOREST

In machine learning, the random forest algorithm is also known as the random forest classifier. It is a widely used classification algorithm. One of the most notable features of this technique is that it can be used for classification as well as random forest regression.

Why Should We Use Random Forest?

Some of the reasons why we utilise the Random Forest Algorithm in machine learning are because of its advantages and key qualities.

- Both classification and regression tasks are suitable for the Random Forest algorithm.
- Cross validation ensures a higher level of accuracy.
- The accuracy of a random forest classifier can be maintained even when missing values are present.
- a significant amount of data.
- When there are more trees in the model, over-fitting trees are not allowed.
- It can deal with a large data set that has a larger dimensionality.

RANDOM FOREST IS USED IN

- Banking Industry
  - Credit Card Fraud Detection
  - Customer Segmentation
  - Predicting Loan Defaults on LendingClub.com's.
- Medicine and healthcare
  - Prediction of Cardiovascular Disease
  - Diabetes Diagnosis
  - Breast Cancer Prognosis
- Stock Market
  - Prediction of the Stock Market
  - Analysis of Stock Market Sentiment
  - Detection of Bitcoin Prices
- E-Commerce
  - Recommendation of a product
  - Price reductions
  - Search Engine Positioning

IMPLEMENTATION

Dataset

To investigate and analyze the recommended strategy of data mining technology for this current challenge, with a focus on patient data.

The value of "class" in the output column is either "1" or "0." The value "0" indicates that the patient does not have CKD, whereas the value "1" indicates that the patient does have CKD.

S. No	Parameters	Non NULL count	Datatype
0	age	391	float64
1	Blood pressure	388	float64
2	Specific gravity	353	float64
3	albumin	354	float64

4	sugar	351	float64
5	Red_blood_cells	248	object
6	Pus_cell	335	object
7	Pus_cell_clumps	396	object
8	bacteria	396	object
9	Bloodglucoserandom	356	float64
10	Blood urea	381	float64
11	Serum creatinine	383	float64
12	sodium	313	float64
13	potassium	312	float64
14	hemoglobin	348	float64
15	Packed_cell_volume	329	object
16	White_blood_cell_count	294	object
17	Red_blood_cell_count	269	object
18	hypertension	398	object
19	Diabetes_mellitus	398	object
20	Coronary_artery_disease	398	object
21	appetite	399	object
22	Peda_edema	399	object
23	anemia	399	object
24	class	400	object

Table: Dataset of CKD

The SMOTE algorithm

Oversampling the minority class is done using the Synthetic Minority Oversampling Technique (SMOTE). A balancer is another name for it. It accepts the entire dataset as input but only works on the minority class. It raises the proportion of people in the minority class. KNN was utilised by SMOTE to find new instances. In the vast majority of situations, it has no effect. The new examples aren't just rehashes of previous minority cases. Instead, the computation performs component space tests for each target class and its closest neighbours, then generates new models that combine the goal case's qualities with the highlights of its neighbours. This method increases the number of features available to each class and broadens the scope of testing.

SMOTE is a data augmentation algorithm that creates synthetic data points depending on the original data points. SMOTE can be thought of as a more advanced variant of oversampling or as a specific data augmentation process. SMOTE has the advantage of not creating duplicate data points, but rather synthetic data points that are somewhat different from the original data points. SMOTE is a better oversampling option.

The following is how the SMOTE algorithm works:

- You select a representative sample from the minority group at random.
- You will identify the k closest neighbours for the observations in this sample.
- The next step is to choose one of those neighbours and determine the vector between the current data point and that neighbour.
- You multiply the vector by an integer between 0 and 1 at random.
- You add this to the current data point to get the synthetic data point.

This technique is essentially the same as moving the data point slightly in the direction of its neighbors. This ensures that your synthetic data point is not an exact duplicate of an existing data point, while simultaneously ensuring that it is not too dissimilar from known observations in your minority class.

RESULT

The influence of imbalanced classes on logistic regression appears to have had a significant impact on the model's specificity, since there are 0 False Positives out of a total of 1,250 predictions in this scenario. Unbalanced classes, in general, do not provide enough exposure to the minority class for models to recognise a minority instance in predictions.

FUTURE WORK

This study paper discussed how to improve the analytical framework for predicting kidney function. As illustrated in Figure 1, the approach adopted in this study follows the basic stages of data mining. To begin, a data collection was acquired and combined to create the target dataset. Traditional techniques for removing missing values are used in the data preparation phase, and a normalisation technique is used to remove bias and defects found in the data. Finally, various rule-based induction models and Decision tree models are used to classify the generated hidden information, allowing it to be interpreted. Decision tree models perform better than rule-based models in terms of accuracy. In the case of imbalanced data, the accuracy recorded is better when SMOTE is used. SMOTE may also be used for Big Data analysis utilizing the Hadoop framework, using a map-reduce programming style with a novel algorithmic approach, which is something we'll be working on in the future.

CONCLUSION

In healthcare data analytics, detecting infrequent but significant healthcare events in vast unstructured datasets is now a regular challenge. This research is the first to try a systematic search for unusual events in unstructured healthcare datasets with an asymmetric distribution. We create a classification framework for healthcare data analytics that includes numerous rebalancing procedures, as well as some tips for dealing with comparable issues. This research has the potential to improve the health-care system and serve as a beneficial tool for doctors in disease prediction. It will also assist clinicians in realising that if a patient can be treated, they can concentrate on the primary risk factors. This research's future work can be done using numerous combinations of machine learning models to take advantage of their combined benefits**.**

REFERENCES

[1] Bishop CM: Pattern Recognition and Machine Learning (Information Science and Statistics). 2007, New York: Springer

[2] He H, Garcia EA: Learning from imbalanced data. IEEE Trans Knowledge Data Eng. 2009, 21 (9): 1263-1284.

[3] Daskalaki S, Kopanas I, Avouris N: Evaluation of classifiers for an uneven class distribution problem. Appl Artif

Intell. 2006, 20 (5): 381-417. 10.1080/08839510500313653.

[4] Ramaswamy S, Ross KN, Lander ES, Golub TR: A molecular signature of metastasis in primary solid tumors. Nat Genet. 2003, 33: 49-54. 10.1038/ng1060.

[5] Shipp MA, Ross KN, Tamayo P, Weng AP, Kutok JL, Aguiar RC, Gaasenbeek M, Angelo M, Reich M, Pinkus GS, Ray TS, Koval MA, Last KW, Norton A, Lister TA, Mesirov J, Neuberg DS, Lander ES, Aster JC: Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. Nat Med. 2002, 8: 68-10.1038/nm0102-68.

[6] Iizuka N, Oka M, Yamada-Okabe H, Nishida M, Maeda Y, Mori N, Takao T, Tamesa T, Tangoku A, Tabuchi H, Hamada K, Nakayama H, Ishitsuka H, Miyamoto T, Hirabayashi A, Uchimura S, HamamotoY:Oligonucleotide microarray for prediction of early intrahepatic recurrence of hepatocellular carcinoma after curative resection. Lancet. 2003, 361 (9361): 923-929. 10.1016/S0140-6736(03)12775-4.

[7] Blagus R, Lusa L: Class prediction for high-dimensional class-imbalanced data. BMC Bioinformatics. 2010, 11: 523+-10.1186/1471-2105-11-523.

[8] Hulse JV, Khoshgoftaar TM, Napolitano A: Experimental perspectives on learning from imbalanced data. Proceedings of the 24th international conference on Machine learning. 2007, Corvallis, Oregon: Oregon State University, 935-942.

[9] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP: SMOTE: synthetic minority over-sampling technique. J Artif Intell Res. 2002, 16: 341-378.

[10] Cieslak DA, Chawla NW, Striegel A: Combating imbalance in network intrusion datasets. Proc IEEE Int Conf Granular Comput. 2006, Atlanta, Georgia, USA, 732-737.

[11] Liu Y, Chawla NV, Harper MP, Shriberg E, Stolcke A: A study in machine learning from imbalanced data for sentence boundary detection in speech. Comput Speech Lang. 2006, 20 (4): 468-494. 10.1016/j.csl.2005.06.002.

[12] Johnson R, Chawla N, Hellmann J: Species distribution modelling and prediction: A class imbalance problem. Conference on Intelligent Data Understanding (CIDU). 2012, 9-16. 10.1109/CIDU.2012.6382186.

[13] Fallahi A, Jafari S: An Expert System for Detection of Breast Cancer Using Data Preprocessing and Bayesian Network. Int J Adv Sci Technol. 2011, 34: 65-70.

[14] Batuwita R, Palade V: microPred: effective classification of pre-miRNAs for human miRNA gene prediction. Bioinformatics. 2009, 25 (8): 989-995. 10.1093/bioinformatics/btp107.

[15] Xiao J, Tang X, Li Y, Fang Z, Ma D, He Y, Li M: Identification of microRNA precursors based on random

forest with network-level representation method of stem-loop structure. BMC Bioinformatics. 2011, 12: 165+-10.1186/1471-2105-12-165.

[16] MacIsaac KD, Gordon DB, Nekludova L, Odom DT, Schreiber J, Gifford DK, Young RA, Fraenkel E: A hypothesis-based approach for identifying the binding specificity of regulatory proteins from chromatin immunoprecipitation data. Bioinformatics. 2006, 22 (4): 423-429. 10.1093/bioinformatics/bti815.

[17] Wang J, Xu M, Wang H, Zhang J: Classification of imbalanced data by using the SMOTE algorithm and locally linear embedding. International Conference on Signal Processing. 2006, Guilin, China

[18] Doyle S, Monaco J, Feldman M, Tomaszewski J, Madabhushi A: An active learning based classification strategy for the minority class problem application to histopathology annotation. BMC Bioinformatics. 2011, 12: 424+-10.1186/1471-2105-12-424.

[19] Wallace B, Small K, Brodley C, Trikalinos T: Class imbalance, Redux. Data Mining (ICDM), 2011 IEEE 11th International Conference on. 2011, Vancouver, Canada, 754-763.

[20] Ertekin SE, Huang J, Bottou L, Giles CL: Learning on the border: Active learning in imbalanced data classification. Proceedings of ACM Conference on Information and Knowledge Management. 2007, Lisbon, Portugal, 127-136.

[21] Radivojac P, Chawla NV, Dunker AK, Obradovic Z: Classification and knowledge discovery in protein databases. J Biomed Inform. 2004, 37 (4): 224-239. 10.1016/j.jbi.2004.07.008.

[22] Han H, Wang WY, Mao BH: Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning. Advances in Intelligent Computing Volume 3644 of Lecture Notes in Computer Science. 2005, Berlin/Heidelberg: Springer, 878-887.

[23] Bunkhumpornpat C, Sinapiromsaran K, Lursinsap C: Safe-Level-SMOTE:Safe-Level-Synthetic Minority Over-Sampling TEchnique for Handling the Class Imbalanced Problem. Advances in Knowledge Discovery and Data Mining, Volume 5476. 2009, Berlin / Heidelberg: Springer, 475-482.

[24] Kaggle, “Chronic Kidney Disease Dataset,” https://www.kaggle.com/abhia1999/chronic-kidney-disease.