G-SMOTE
Oversampling for Insurance Data

Fernando Bacao[1*], Georgios Douzas[1]

[1]NOVA Information Management School, Universidade Nova de Lisboa

*Corresponding Author: bacao@novaims.unl.pt

Postal Address: NOVA Information Management School, Campus de Campolide, 1070-312 Lisboa, Portugal

Telephone: +351 21 382 8610

Traditional supervised machine learning classifiers are challenged to learn highly skewed data distributions as they are designed to expect classes to equally contribute to the minimization of the classifiers cost function. Moreover, the classifiers design expects equal misclassification costs, causing a bias for underrepresented classes. Thus, different strategies to handle the issue are proposed by researchers. The modification of the data set has become a common practice since the procedure is generalizable to all classifiers. Various algorithms to rebalance the data distribution through the creation of synthetic instances were proposed in the past. In this paper, we propose a new oversampling algorithm named G-SOMO, a method that is adopted from our previous research. The algorithm identifies optimal areas to create artificial data instances in an informed manner and utilizes a geometric region during the data generation to increase variability and to avoid correlation. In this paper we use real world data from a Columbian insurance company, to investigate the benefits of using oversampling to improve the quality of predictive machine learning models. Our empirical results, validated with different classifiers and metrics against a benchmark of commonly used oversampling methods show that G-SOMO consistently outperforms competing oversampling methods.

1 Introduction

In recent years machine learning models have become pervasive in most economic activities and human endeavors. The growing popularity of these models comes mostly from the raising of two different but related trends: cheap computation and huge amounts of data. This binomial created the opportunity to be more accurate in problems where there is very limited theoretical knowledge. Basically, machine learning harnesses the power of data and computing to produce more accurate predictions. The use of machine learning models to assess and predict different types of risks has become a common practice in most insurance companies around the world, as these models have proved to be able to increased accuracy in most risk related predictions tasks.

Being data a fundamental tenant of machine learning, its quality and availability will obviously impact the quality of the final models. In many situations the availability of data becomes the main problem in building accurate and robust machine learning models. Although, in general we can say that we live in a data-rich world there are still many domains and circumstances characterized by data scarcity. In risk analysis a particular type of data scarcity tends to be prevalent, resulting from the fact that it deals

with uncommon events; as a consequence datasets are usually highly skewed. Examples of this range from assessment of financial credit risk to fraud detection, and include forecasting future high-cost users of health care and bankruptcy prediction models. It is safe to say that in order to successfully apply machine learning algorithms to most risk assessment problems the user will be faced with the need to solve, or at least mitigate, the negative effects of dataset imbalance.

Building models with highly skewed datasets is usually known, in the machine learning literature, as imbalanced learning. Imbalance learning is a pervasive issue in risk analysis and assessment, as finding perfectly balanced datasets, with 50% examples of each class, is very rare. Imbalance learning constitutes asignificant obstacle when building accurate and robust predictive models. Skewed datasets present an imbalance between the different classes represented. In these cases, the datasets may even be very large, but they are characterized by having a large number of data points of one class (majority class) and a very small number of data points of the other class(es) (minority classes). This is not an optimal situation, as the objective of a classification algorithm is to learn a classifier that is able to discriminate the classes [He and Garcia, 2009]. Having a smallnumber of data points from one of the classes will make the learning task much more challenging, eventually impossible. The ability to deal and mitigate the imbalance learning problem constitutes an important skill for both practitioners and researchers alike.

There are several options to deal with imbalanced learning, which can be broadly classified into three main groups [Fernández et al., 2013]. One is the modification or creation of algorithms that reinforce the learning towards the minority class. Another approach relates with the application of cost-sensitive methods to minimize higher cost errors. The last one consists in the modification of the data, by re-balancing the class distribution, which is done through the use of undersampling, oversampling or hybrid methods. This last approach has some relevant advantages over the previous ones, especially in terms of real world applications, where most users are usually not machine learning experts, and thus not able to modify algorithms. Using dataset re-balancing strategies simplifies the task and, once the dataset is balanced, the user can use any "off-the-shelf" algorithm.

It is important to note that there are two different problems associated with imbalanced learning. The first concerns the skewness of the distribution, in other words, having a relative small number of positive samples when compared with the number of negative samples. In this case we have what can be called relative scarcity. The second relates to the case in which, besides this imbalance, we also have a small number of positive samples in absolute terms. While in the first case we may consider different strategies to approach the problem, in the second case the only real option is oversampling. This paper focuses on improvement of predictive machine learning models through the use of oversampling techniques to deal with the imbalanced learning problem in the context fraud identification. Identifying fraudulent behavior is relevant for financial companies due to the disastrous consequences it brings with it. These usually translate into economic losses, negative impact on its public image, client's forfeiture, among others. In Colombia, this situation constitutes an important problem for insurance companies, as they are faced with significant levels of fraud in compulsory auto insurance. Millions of dollars in losses force companies to search for better methodologies to assertively respond to this state of affairs, and better predictive models are amongst the most promising tools to control this problem. In this paper we use real world data from a Columbian insurance company, to investigate the benefits of using oversampling to improve the quality of predictive machine learning models.

We will compare several oversampling methods, which are readily available for use by practitioners and researchers in the risk analysis domain. The objective is to understand the impact of these oversampling methods in the quality of the machine learning models produced, understand which of the methods is the best option and derive some practical guidelines for its application.

The rest of this paper is structured as follows: in section 3 a general review related with oversampling techniques and imbalanced learning problem is presented. In section 4 research methodology is explained followed by results and discussion in section 5. Finally, conclusions and future work are expressed on

section 6.

2 State of the art

Although companies' efforts in fighting fraud have increased in recent years and more sophisticated methodologies have been used, fraud consequences are still evident in economic losses. According to the (2016 Global Economic Crime Survey, 2016) in the insurance industry these numbers approach 80 billion losses per year across all lines of insurance in the U.S. (Coalition Against Insurance Fraud, n.d.). In Germany the insurance association (GDV) estimates (Hartley, 2016) In Colombia, the National Institute for Investigation against Fraud (INIF) ensures that (Chacón, 2017). It is important to note that fraud repercussions transcend economic losses, also having negative impacts in savings losses, increased premium's for users, distrust in the insurance sector and insurer's image discredit. Fraud, has also produced the need for the insurance sector to improve business processes control, which frequently translate into significant overhead burden and tedious operational tasks, which consume substantial resources, both in terms of budget and time. For these reasons, the industry has invested significantly in the development of machine learning models, hoping to improve the ability to proactively detect and prevent fraudulent behaviors [**?**]. In fact, and given its negative impacts, preventing fraud and reducing its financial impact on the business can be the right strategy to increase competitiveness and grow the bottom line. A significant body of research can be found on the topic of fraud prevention [Phua et al., 2004] [Wei et al., 2012] [Bănărescu, 2015] [Hassan and Abraham, 2015] [Kim et al., 2016] [Sahin et al., 2013]. Most of the research efforts have been focused on finding the best models to identify fraudulent transactions or events, in order to expedite the operational procedures, saving time and money. The Compulsory Auto Insurance (SOAT according to the Spanish acronym) was established in Colombia in 1986, as a response to the increase in automotive accidents. With a total of 1.3 million of fatal victims and more than 50 million injured, this problem can be considered a public health problem. SOAT's objective is to guarantee basic medical treatment to all the victims involved in this type of accident in the country. Unfortunately, because of the mismanagement carried out by the government, health entities and users, many of the cases are, in some way, fraudulent, and the impossibility to determine the premium according to the risk, lead the insurance companies, which provide this coverage, to an state annual losses of more than 54 million dollars. The SOAT is highly susceptible to fraud; the actions made by fraudsters can be enumerated as follows:

- Bills of services which cannot be proved;
- Charging the same bill to different insurance companies;
- Using the policy to charge medical treatments resulting from different situations than car accidents;
- Health companies can charge bills related to nonexistent patients or apply unnecessary treatments like surgeries to the injured person;
- Nonexistent health entities charge fake bills to the insurance companies;
- Inflation of treatments and medicines costs.

Fraud is, by definition, an unusual event, in the insurance industry fraudulent claims are significantly fewer than non-fraudulent cases. Also, the cost of misclassified a fraudulent claim is higher than a considered rightful claim as a counterfeit claim. Hence, the development of machine learning models for the identification of fraudulent claims in the insurance industry can be defined as an imbalanced learning problem. Most machine learning models perform poorely in imbalanced data: they can have a good accuracy for majority (negative) class but poor results for the class of interest (positive) [He and

Garcia, 2009]. Additionally, differentiate between the minority class and noise is challenging: examples belonging to the minority class can be identified by the algorithm as noise, and in the same way noisy individuals can be treated as the individuals of interest [Beyan and Fisher, 2015] The relevance of this problem can be seen by the 527 scientific papers published in the last 10 years, with a significant increase in the last few years [Haixiang et al., 2017]. There exist four main groups were the proposed solutions can be categorized:

References

[Bănărescu, 2015] Bănărescu, A. (2015). Detecting and preventing fraud with data analytics. *Procedia Economics and Finance*, 32:1827–1836.

[Beyan and Fisher, 2015] Beyan, C. and Fisher, R. (2015). Classifying imbalanced data sets using similarity based hierarchical decomposition. *Pattern Recognition*, 48(5):1653–1672.

[Fernández et al., 2013] Fernández, A., López, V., Galar, M., del Jesus, M. J., and Herrera, F. (2013). Analysing the classification of imbalanced data-sets with multiple classes: Binarization techniques and ad-hoc approaches. *Knowledge-Based Systems*, 42:97–110.

[Haixiang et al., 2017] Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., and Bing, G. (2017). Learning from class-imbalanced data: Review of methods and applications. *Expert Systems with Applications*, 73:220–239.

[Hassan and Abraham, 2015] Hassan, A. K. I. and Abraham, A. (2015). Modeling insurance fraud detection using imbalanced data classification. In *Advances in Intelligent Systems and Computing*, pages 117–127. Springer International Publishing.

[He and Garcia, 2009] He, H. and Garcia, E. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284.

[Kim et al., 2016] Kim, Y. J., Baik, B., and Cho, S. (2016). Detecting financial misstatements with fraud intention using multi-class cost-sensitive learning. *Expert Systems with Applications*, 62:32–43.

[Phua et al., 2004] Phua, C., Alahakoon, D., and Lee, V. (2004). Minority report in fraud detection. *ACM SIGKDD Explorations Newsletter*, 6(1):50.

[Sahin et al., 2013] Sahin, Y., Bulkan, S., and Duman, E. (2013). A cost-sensitive decision tree approach for fraud detection. *Expert Systems with Applications*, 40(15):5916–5923.

[Wei et al., 2012] Wei, W., Li, J., Cao, L., Ou, Y., and Chen, J. (2012). Effective detection of sophisticated online banking fraud on extremely imbalanced data. *World Wide Web*, 16(4):449–475.